# The landscape, properties, and determinants of transcriptional activation of endogenous transposable elements in grapevine (*Vitis vinifera* L.)

---

A thesis

submitted in partial fulfilment

of the requirements for the Degree of

Doctor of Philosophy

at

Lincoln University

by

Ting-Hsuan Chen

---

Lincoln University

2020

Abstract of a thesis submitted in partial fulfilment of the requirements for the Degree of Doctor of Philosophy.

# The landscape, properties, and determinants of transcriptional activation of endogenous transposable elements in grapevine (*Vitis vinifera* L.)

by

Ting-Hsuan Chen

Transposable elements (TEs) are an intrinsic mutagen of eukaryotic genomes and have been proposed to be important in increasing genetic diversity in plants. It has been known that biotic and abiotic stress treatments induce TE transcription, the first stage in TE mobilisation. This research began with an investigation of TE transcription activity in grapevine embryogenic callus subjected to biotic stressors (*Botrytis cinerea* extracts and live *Hanseniaspora uvarum* cultures) to determine the location and regulation of autonomous TEs.

Short-read RNA sequencing (RNAseq) has been commonly used to determine TE transcription patterns at a family level. This research sought to further these approaches by establishing an analysis pipeline to identify the expression of individual TE loci from Illumina RNAseq data. We efficiently identified that only 1.7%-2.5% of total annotated TE loci were transcribed in our system. This work identified a strong tendency for TE expression candidates to be found within introns of expressed genes. It was also discovered that these pairs of TEs and genes shared the same differential expression patterns in response to applied stressors.

Our analysis pipeline was successfully validated using publically available RNAseq datasets from *Arabidopsis,* wild-type and epigenetic mutant (*ibm2* and *ddm1*) lines, and *Drosophila* datasets of amyotrophic lateral sclerosis (ALS) models exhibiting a TE transcriptional storm. We successfully identified an *Arabidopsis COPIA-93* locus previously proven to mobilise in *ddm1* mutant and a subset of *Drosophila* TE loci that potentially contributed to full-length autonomous TE transcripts in the ALS models that have not been previously reported.

Oxford Nanopore Technology (ONT) cDNA sequencing was deployed to determine whether autonomous TEs were being expressed as a precursor of mobilisation. Only low levels of full-length

transcription of one Gypsy-V1 locus and three hAT-7 loci was detected in this data, suggesting rare intact transcription from autonomous TE loci despite stress treatments. This finding suggested that TE mobilisation might require inhibition of the epigenetic silencing system.

We, therefore, treated embryogenic callus with the histone deacetylase inhibitors (HDACi), trichostatin A (TSA) or 4-phenylbutyric acid (4PBA), to alter the heterochromatic architecture of callus cells. Only the 4PBA treatment showed a noticeable shift in the transcriptional landscape of TE transcription, significantly increasing the proportion of intergenic TE loci in the expression candidate pool and resulting in significant up-regulation of 2,059 TE loci. ONT cDNA sequencing of these samples detected very low levels of intact sequencing reads from different yet a single Gypsy-V1 locus and six hAT-7 loci. Five genes participating in the RNA-dependent DNA methylation (RdDM) pathway (AGO2, AGO4, RDR1, RDR6, and NERD) were upregulated, suggesting that callus exposed to 4PBA responded by an enhancement of RdDM, maintaining effective control of TE transcription and therefore TE mobility.

Overall, this thesis contributes to the understanding of the landscape, properties, and determinants of transcriptional activation of endogenous transposable elements, revealing the closely connected transcriptional relationship between TEs and co-localised genes. These findings shed light on the genetic and epigenetic impact of endogenous TE activation on genes in nature.

**Keywords:** transposable element, transcription, epigenetic silencing, grapevine, stress treatment, histone deacetylase inhibitor, *Vitis vinifera*

# Acknowledgements

Research is an expedition that explores the unknown area of science, and PhD training is just the beginning of this expedition. During the journey of pursuing my PhD, there are difficulties, challenges, anxiety, and self-doubt. I'm truly grateful that many people have helped and encouraged me throughout all these processes, so I have the opportunity to enjoy the pleasure of doing science.

Thank you to my supervisor, Assoc. Prof. Chris Winefield, for your generous investment of intelligence and time in guiding me to do research. Your perseverance and enthusiasm in science have inspired me to set high standards and to be self-critical in the research work. Thank you for encouraging me to aim high and push the boundary of science. It has been a pleasure to be your mentee and colleague, and I look forward to working with you for many years to come.

I am immensely thankful to my associate supervisor, Assoc. Prof. Stephen On, for sharing his wisdom of being a responsible and successful scientist. I would also like to express my gracious appreciation to my co-supervisor, Dr Ross Bicknell (Programme Leader at Plant and Food Research: Grape Genetics), and the members of his research team in Plant and Food Research: Dr Philippa Barrell, Dr Susan Thomson, Tim Miller, and Michelle Thompson, for their valuable friendship, inspiration, and experimental support.

During these years of PhD study, my computational coding and bioinformatic analysis skills have grown from zero to being able to establish analysis workflows for this research independently. All this progress started from a 6-page script from my advisor, Dr Darrell Lizamore (now Principal Research Scientist at Bragato Research Institute). Darrell has inspired me to think like a bioinformatician and think out of the box. He has my sincere gratitude.

I am most grateful to have been supported by the New Zealand International Doctoral Research Scholarship and appreciate that this research work has been funded by the Grape Genetic Improvement Project of Plant and Food Research Strategic Science Investment Fund.

I have a deep thankfulness for the kindness and friendship of my colleagues in the Department of Wine, Food, and Molecular Biosciences during these past few years. I especially thank Lorraine Holmes, our Department Secretary, who has been looking out for students of this department and saved me from being locked out from the office after working hours when I mistakenly left all my keys (including car and home keys) in the office. I also profoundly thank our lab manager, Dr Pani Vijayan, who joined our team in late 2019, for her friendship and sense of humour that has been a morale boost during the stressful period of writing the thesis.

Last but not least, I am indebted to my family for their love, encouragement and support. Thank you to my husband, Yen-Heng, for your thoughtfulness, patience and ability to cheer me up. I am blessed and fortunate to have you with me on this journey (of life).

# Table of Contents

# List of Tables

*Appendices*

# List of Figures

*Appendices*

# Acronyms

| | |
|---|---|
| Alt3 | Alternative 3’ splicing |
| Alt5 | Alternative 5’ splicing |
| AFLP | Amplified fragment length |
| AGO | Argonaute protein |
| AtHDA | Arabidopsis histone deacetylase |
| bp | Base pair(s) |
| CaMV | *Cauliflower mosaic virus* |
| CAP | Capsid-like protein |
| cDNA | Complementary DNA |
| CDS | Coding DNA sequence |
| C-flank | C-terminal 2kb-flanking region of a gene |
| CMT | Chromomethylase |
| CRE | cis-regulatory element |
| DBD | DNA binding domain |
| DCL | Dicer-like ribonuclease III family |
| DDM1 | Decrease in DNA methylation 1 (chromatin-remodeling protein) |
| DEG | Differentially expressed gene |
| DETE | Differentially expressed transposable element |
| DME | Demeter DNA glycosylase |
| DML | Demeter-like DNA glycosylase |
| DNA | Deoxyribose nucleic acid |
| DRM2 | Rearranged Methyltransferase 2 (a DNA methyltransferase) |
| dsRNA | Double-stranded RNA |
| EDM2 | Enhanced downy mildew 2 protein |
| epiRIL | Epigenetic recombinant inbred line |
| ER | Endoplasmic reticulum |
| ES | Exon skipping |
| EVD | Evadé retrotransposon |
| FPKM | Fragments per kilobase of exon model per million reads mapped |
| Gb | Gigabase pairs |
| GFP | Green fluorenscent proetin |
| Gly | Glycine |
| GO | Gene ontology |
| GUS | Beta-galactosidase |
| H3K4me | Methylation at lysine 4 of histone H3 |
| H3K9me | Methylation at lysine 9 of histone H3 |
| H3K27me | Methylation at lysine 27 of histone H3 |
| H3K36me | Methylation at lysine 36 of histone H3 |
| HAT | Histone acetylase |
| hAT | hobo, Activator and Tam3 transposon |
| hATC | hAT C-terminal dimerization |
| HDAC | Histone deacetylase |
| HDACi | Histone deacetylase inhibitor |
| IBM1 | Increase in BONSAI methylation 1 (a Jumonji C family histone demethylase) |
| IBM2 | Increase in BONSAI methylation 2 (an RNA binding protein) |
| IN | Integrase |
| INT | Internal domain |
| IR | Intron retention |

| | |
|---|---|
| JMJ14 | Jumonji C domain-containing protein 14 (a histone demethylase) |
| kb | Kilobase pairs |
| KYP | Kryptonite |
| LDL | Lysine-Specific Demethylase 1-Like |
| LINE | Long interspersed repetitive element |
| LTR | Long-terminal repeat |
| LTR-TE | LTR transposable element |
| Mb | Megabase pairs |
| MET | Methyltransferase 1 |
| miRNA | Micro RNA |
| MITE | Miniature inverted-repeat transposable element |
| mRNA | Messenger RNA |
| MULE | Mutator-like element |
| NERD | Needed for RDR2-independent DNA methylation (an AGO binding protein) |
| N-flank | N-terminal 2kb-flanking region of a gene |
| NGO | No start codon for mRNA translation |
| NST | Having start codon but no stop codon for mRNA translation |
| nt | Nucleotide |
| ONT | Oxford nanopore technology |
| ORF | Open reading frame |
| 4PBA | 4-phenylbutyric acid |
| PBS | Primer binding site |
| Pol II | RNA polymerase II |
| Pol IV | RNA polymerase IV |
| Pol V | RNA polymerase V |
| PR | Proteinase |
| PRO | Productive in terms of mRNA translation into protein |
| PPT | Polypurine tract |
| PTC | Premature termination codon for mRNA translation |
| PTGS | Post-transcriptional gene silencing |
| RdDM | RNA dependent DNA methylation |
| RDR | RNA-dependant RNA polymerase |
| RH | Rnase H |
| RNA | Ribose nucleic acid |
| RNAi | RNA interference |
| RNAseq | RNA sequencing |
| ROS1 | Repressor of silencing 1 DNA glycosylase |
| RPM | Reads per million mapped reads |
| rRNA | Ribosomal RNA |
| RT | Reverse transcriptase |
| SDE3 | Silencing defective protein 3 (an RNA helicase) |
| SINE | Short interspersed repetitive element |
| siRNA | Small interference RNA |
| sRNA | Small RNA |
| ssDNA | Single-stranded DNA |
| SUVH | Su(var)3-9 homolog |
| SWIM | SWI2/SNF2 and MuDR |
| TE | Transposable element |
| tasiRNA | Trans-acting silencing RNA |
| TIR | Terminal inverted repeat |
| TIR-TE | TIR transposable element |

| | |
|---|---|
| tRNA | Transfer RNA |
| tRF | tRNA fragment |
| TSA | Trichostatin A |
| TSS | Transcriptional start site |
| TPase | Transposase |
| UTR | Untranslated region |
| UV | Ultraviolet |
| VLP | Virus-like particle |

# Chapter 1

# Introduction

## 1.1 Discovery and significance of transposons

In the late 1940s, Barbara McClintock challenged the existing concept of gene properties by discovering a DNA segment that could change its location in maize chromosome 9 (McClintock, 1950; Ravindran, 2012). This locus was named "*Ds*" for its property to 'dissociate' from its original chromosomal location. Her studies found that *Ds* could only mobilise in the presence of another locus, named *Activator* or *Ac*, that could transpose independently. McClintock published the first research article describing the mutable loci in 1950 (McClintock, 1950). The year after, she presented her work in discovering *Ac* and *Ds* at a seminal Cold Spring Harbor Symposium. When she finished her presentation, what she gained was not a burst of applause or excited discussion but dead silence (Ravindran, 2012). It was not until the 1970s when molecular biologists discovered mobile DNA loci in other organisms, viruses and bacteria, that the significance of her groundbreaking research work began to be recognized and appreciated (Ravindran, 2012). In 1976, 26 years after McClintock's PNAS classic article was published, the term "transposable elements" (TEs) was officially introduced at a Cold Spring Harbor meeting on "DNA Insertion Elements, Plasmids, and Episomes" as an explicit acknowledgement of her contribution (Keller, 1983). On 10 December 1983, at the Concert Hall in Stockholm, Sweden, she was awarded the Nobel Prize in Physiology or Medicine for her discovery of mobile genetic elements (NobelPrize.org).

During the past 70 years, TEs have been widely found in eukaryotic and some prokaryotic organisms. These constituent components of genomes have surprised us by comprising about two-third of our own genome and 85% of maize's (Fedoroff, 2012). TEs have been frequently considered parasitic or even harmful to the cells due to the self-proliferating and mutagenic characteristics. TE transposition into coding genes can disrupt gene function or cause aberrant alternative splicing (Chuong et al., 2017). However, their existence defines some chromosomal compartmentations, such as sub-telomere and peri-centromeric regions (Slotkin and Martienssen, 2007). Furthermore, a growing weight of evidence shows that TE mobilization has benefited host cells by contributing TE component protein domains to shape gene function or by providing *cis*-regulatory elements that render the co-localized genes responsive to an array of external stimuli, including environmental signals, to alter transcriptional programmes accordingly (Casacuberta and González, 2013; Chuong et al., 2017; Feschotte, 2008; Lisch, 2013).

Despite the possible devastating consequences of a burst of TE transposition, the substantial volume of TEs in genomes of various species across all three kingdoms of life indicates their ancient colonization of these same genomes. What we observe today through the expanding lens of modern structural genomics is the outcome of an ancient arms race between TE activity and the epigenetic silencing machinery, a multi-layered endogenous silencing system regulating gene or TE activity through DNA and histone modifications as well as small RNA mediated interference mechanisms. Genetic drift and natural selection have actively shaped the TE and epigenetic landscapes. With the expansion of 2$^{nd}$ and 3$^{rd}$ generation sequencing approaches, we are now finding that the 'parasite' and 'junk' stereotypes applied to endogenous TEs are only one of the facets of their role in host cells. During their long evolutionary journey, TE segments have been frequently adopted as functional domains or *cis*-regulatory elements for coding genes that have consequently gained tissue-specific properties, developmental-stage specificities, and often stress responsiveness. Retention of such new genetic variation, along with host initiated epigenetic silencing strategies acting to silence active elements, is likely to have been affected upon by natural selection, resulting in enhanced survival or reproductive fitness (Baduel et al., 2019; Casacuberta and González, 2013; Chuong et al., 2017; Feschotte, 2008; Rech et al., 2019; Rey et al., 2016). This involvement of TEs in modulating gene functionality sometimes makes particular TE segments necessary for the host cells, preventing them from being removed from the host genome through evolutionary time.

Crop domestication is one of the underpinning activities that has driven the development of human civilization. Key domestication phenotypes, such as hermaphroditic flowering, self-compatibility, increased yield, improved flavour, altered harvest timing, have been explicitly selected for and are often monogenic traits. This strong selection often leads to a rapid reduction in the genetic diversity of domesticated crops. However, for crop plants that are easily vegetatively cloned (typically long-lived perennial crops like fruit trees and crops with complex genetics such as potatoes and orchids), there exists a critical duality where desirable phenotypes are immortalised at an individual level, but further evolution of the population is frozen in time. This ultimate form of genetic fixation leaves these crops exposed to ever-changing abiotic and evolving biotic stress, without an apparent mechanism to evolve or an easily breed adaptive strategy.

In recent decades, various tools and methods have been developed to expand the possibilities for breeding new crop varieties (Lusser et al., 2012; Springer and Schmitz, 2017; Thieme and Bucher, 2018). While random point mutations and large insertions or deletions (INDELs) can be generated by chemical and physical mutagens, site-directed mutagenesis of specific genes can be achieved by various technologies, but this would inevitably introduce artificial or exogenous DNA into crop genomes (Lusser et al., 2012). Contrarily, mobilization of endogenous TEs is part of the intrinsic

nature of eukaryotic organisms. Therefore utilization of the mutagenic characteristic of endogenous TEs could be one of the long-term solutions to agriculture sustainability (Thieme and Bucher, 2018).

TEs often remain active at very low levels in somatic tissues, as evidenced by the accumulation of new TE insertions in prolonged plant tissue culture and the resulting plant regenerants (Hirochika, 1993; Hirochika et al., 1996; Miyao et al., 2003; Peschke et al., 1987; Yamashita and Tahara, 2006), as well as by the contribution of TEs to the relatively regular appearance of bud sports (Foster and Aranzana, 2018) and the polymorphism of vegetatively propagated clones (Carrier et al., 2012). These observations highlight the potential of TEs to generate dominant or semi-dominant mutations that lead to phenotypes of considerable interest in agriculture.

Many economically and culturally important grapevine varietals have been clonally propagated for long periods, while the development of new varieties has heavily relied on bud sports to generate new phenotypic variation (Pelsy, 2010). This reliance on clonal production and the spread of growing areas to the majority of land masses have exposed this crop to many biotic and abiotic challenges that it is ill-prepared to combat due in the main to contracted effective population size (Charlesworth, 2009; Liang et al., 2019). Simulations by Hannah et al. (Hannah et al., 2013) show that the areas suitable for viticulture will decrease 19% to 73% in major grape-producing regions by 2050. Extreme weather events and regional climatic variation outside what is considered normal bounds elevate the risks to disease, pathogens and pests (Malheiro et al., 2010; Nesbitt et al., 2016; Sturman and Quénol, 2012). Warmer areas are generally wetter and favour fungal diseases such as powdery and downy mildew, whereas cooler regions suffer from frosts. Since 1941, Marlborough, NEW Zealand, has shown a trend of increasing temperature variation with widening variation in the extremes of temperature (Sturman and Quénol, 2012), a great concern to grape growers. Without the intervention of adaptive genetic improvement, alternating strategies of viticultural practices and management with existing cultivars is currently the only available solution to industry (Santos et al., 2020). These studies emphasize the need of establishing a platform to develop populations with increased genetic variation yet retaining the historical pedigree that is strongly favoured by the market. Such a technical and genetic platform has been pioneered by the Winefield group utilising endogenous TEs to produce new clonal material efficiently (Lizamore, 2013). Increased transcriptional activity of a subset of TEs has been observed in grapevine embryogenic callus that was subjected to a range of biotic stress treatments in addition to wound-like pre-treatment, accompanying with an array of phenotypic variation in grapevines regenerated from these treatments (Lizamore, 2013). However, the identification of individual active TE loci and the responses of genes and epigenetic networks have remained unclear. By investigating the characteristics of transcriptionally active TEs, the factors important for TE activation and the

associated epigenetic responses to these transcriptional changes will be determined to uncover how to trigger autonomous TE mobilization in a 'wild-type' Pinot noir background.

Mutants impaired in DNA methylation, histone modification, chromatin remodelling, or small RNA biogenesis have exposed the role of epigenetic responses in regulating TE activity and mobility. In these mutants, a failure of part of the epigenetic regulatory network leads to the heightened and continuous TE activity (Ito et al., 2011; Marí-Ordóñez et al., 2013; Yu et al., 2017; Zemach et al., 2013), which can be further enhanced under conditions of stress (Cavrak et al., 2014; Ito et al., 2011, 2016). The combination of these multiple epigenetic mechanisms in restraining TE activity is remarkable in its effectiveness. Identification of the 'natural' triggers (i.e. stressors) and epigenetic conditions facilitating heightened TE mobility in 'wild-type' plant is a complicated and multifaceted challenge. Yet dissecting these regulatory mechanisms is necessary to artificially enhance TE mobilisation rates to the levels that efficiently impact phenotypic variation in breeding populations. Equally, these insights provide an opportunity to recapitulate a historic TE burst, which would enable the assessment of such transposition bursts on genome function. From the genetic effects made by TE insertions to the consequential epigenetic reprograming against the active TEs, this re-stimulated transposition of endogenous TEs might shed light on the host's exaptation process of both the genetic and epigenetic modifications at a population scale.

## 1.2 Types of TEs

Transposable elements are categorized into two groups based on their dependence on RNA intermediates during transposition. Each type is organised hierarchically into subgroups as follows (Finnegan, 1992):

### 1.2.1 Type I TEs: retrotransposons

Type I TEs are also known as retrotransposons and retro-elements (Lisch, 2013), which transpose through an RNA intermediate-dependent process in a 'copy-and-paste' manner. Depending on the presence or absence of the long-terminal repeats (LTR) flanking the core sequences, Type I TEs can be further grouped into LTR retrotransposons or non-LTR retrotransposons (Slotkin and Martienssen, 2007).

#### 1.2.1.1 LTR retrotransposons (LTR-TEs)

The identical pair of LTRs appeared at both ends of an LTR-TE is a key feature of this group. The 5' LTR supplies the transcription initiation signal to synthesise full-length LTR-TE mRNA, and the 3' LTR provide the transcription termination and polyadenylation signals (Schulman, 2013). The 3' end of the 5' LTR is attached with the primer-binding site (PBS) for the initiation of the first strand reverse transcription, while the 5' end of the 3' LTR is linked with the polypurine tract (PPT), serving as the priming site for the second strand cDNA synthesis (Schulman, 2013). As shown in Figure 1.1 A, the core DNA sequences of LTR-TEs comprise *gag* gene encoding a capsid-like protein (CAP) and *pol* gene encoding integrase (IN), reverse transcriptase (RT), RNaseH (RH) as well as proteinase (PR; Slotkin and Martienssen, 2007; Wicker et al., 2007). Notably, the proteins encoded by the *pol* gene are arranged in a different order in the two major superfamilies *Copia* and *Gypsy* (Schulman, 2013).

To mobilize in a host genome, mRNAs of LTR-TEs transcribed from RNA polymerase II (Pol II) serve as templates for both the translation of the core proteins encoded by *pol* and *gag* and the reverse transcription by RT into cDNAs, which are inserted into new positions in genomic DNA through the activity of IN, leading to an increase in genome size.

#### 1.2.1.2 Non-LTR retrotransposons (non-LTR-TEs)

Non-LTR retrotransposons are subdivided into long interspersed repetitive elements (LINEs) and short interspersed repetitive elements (SINEs; Wessler, 2006). The classical model of LINE sequence structure in plants comprises the 5' untranslated region (UTR) containing transcription initiation site, the 3' UTR annealed with a polyA tail, and two open reading frames, ORF1 and ORF2 (Figure 1.1 B; Schulman, 2013). It is believed that ORF1 encodes a protein with RNA binding and nucleic acid

chaperone properties essential for the self-competent proliferation of LINEs, whereas ORF2 provides the enzymatic activities (PR, RT and RNase H) for cDNA synthesis (Martin, 2010; Schulman, 2013).

Although SINEs are incompetent of self-proliferation and dependent on the enzymes of LINEs for mobilization, they are not derivatives of autonomous LINEs (Wicker et al., 2007). The Pol III promoter harboured in the 5' head of SINEs reveals their origin from tRNA, 5S rRNA and the 7SL signal recognition particle RNA (Wicker et al., 2007). Despite the fact that the discovery of SINEs in many plant families, such as *Brassicaceae*, *Commelinaceae*, *Fabaceae*, *Gramineae*, *Rosaceae*, and *Solanaceae* (Deragon and Zhang, 2006), their presence in *Vitaceae* genomes is not reported yet.

### 1.2.2 Type II TEs: DNA transposons

Most type II TEs are DNA transposons that mobilize in a process described as 'cut-and-paste' transposition. The terminal inverted repeats (TIRs) that bound these elements are recognized by the transposase (TPase) encoded by the same type of TEs (usually TEs of the same family), triggering excision of both strands of DNA at each end followed by the reintegration of DNA transposons from donor sites to the new recipient site in the host genome. DNA transposons generally mobilize without increasing copy number. However, the gap at the donor site created by the double-stranded break after excision can be either sealed without element replacement or repaired by filling with a copy of the transposon, as the latter results in duplicative transposition (Slotkin and Martienssen, 2007).

In *Vitis vinifera*, *hAT* (*hobo*, *Activator* and *Tam3*) and *Mutator*-like element (MULE) are the two superfamilies of type II TIR transposon (TIR-TE) possessing the classical structure comprised of TIRs and a single ORF encoding TPase, which usually contains a DNA binding domain (DBD) and a TPase core enzymatic domain (Figure 1.1 C; Benjak et al., 2008; Feschotte, 2008; Lizamore, 2013). The other two TIR-TE superfamilies, *Harbinger* and CACTA, are also reported in the grapevine genome (Benjak et al., 2008; Lizamore, 2013). Unlike *hAT* and MULE, their DBD and TPase core domains are encoded by two distinct ORFs (Buchmann et al., 2014; Feschotte, 2008; Frey et al., 1990; Grzebelus et al., 2007; Jiang et al., 2003).

*Helitron* is a novel category of Type II TEs that has been proposed, primarily based on *in silico* study, to mobilize using a 'rolling circle' mechanism, in which a single-stranded DNA (ssDNA) of *Heliton* excised from the donor site of the host genome forms a circular ssDNA intermediate before integrating into a different recipient site (Grabundzija et al., 2016; Kapitonov and Jurka, 2007; Slotkin and Martienssen, 2007). The completion of *Helitron* proliferation involves DNA replication at the donor and the recipient sites (Grabundzija et al., 2016; Kapitonov and Jurka, 2007). *Helitron* elements were initially identified by *in silico* analysis of genomic sequences of *Arabidopsis thaliana*, *Oryza*

*sativa*, and *Caenorhabditis elegans*, and then widely discovered in various eukaryotic species, where it is suggested to encode an enzyme comprised of replicase and helicase domains that are necessary for its autonomous mobilization (Kapitonov and Jurka, 2007). Active *Helitron* elements, however, have not yet identified nor isolated from any species, hence the lack of evidence for their transposition mechanism (Grabundzija et al., 2016).

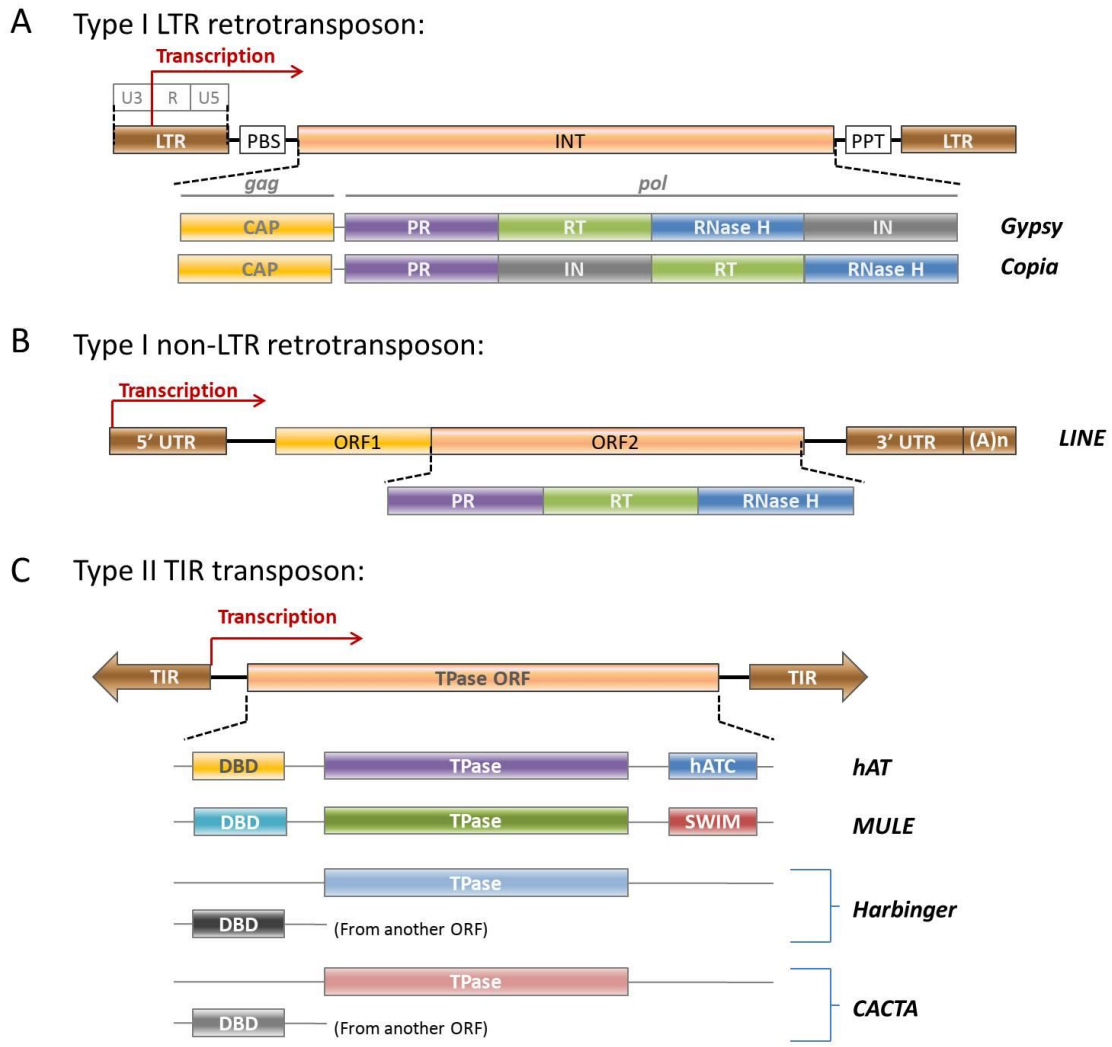



**Figure 1.1 Types and structures of common TEs in grapevines**

**(A)** Key features of type I LTR retrotransposons (LTR-TEs) include LTRs (long-terminal repeats) comprised by U5 (unique 5' region), R (repeated region), and U3 (unique 3' region), PBS (primer binding site), PPT (polypurine tract), and INT (internal domain) containing *gag* gene encoding CAP (capsid-like-protein) and *pol* gene encoding PR (protease), RT (reverse transcriptase), RNase H and IN (integrase). **(B)** The major superfamily of autonomous type I non-LTR retrotransposons (non-LTR-TEs), LINE, is constituted of 5' UTR (untranslated region) containing transcription start site, 3' UTR attached with polyadenylation signal, and two ORFs (open reading frames), one of which encodes PR, RT and RNase H. **(C)** type II TIR transposons (TIR-TEs) are often characterised with terminal inverted repeats (TIRs) flanking an ORF that encodes transposase comprising putative DBD (DNA binding domain), transposase core active domain (TPase) and superfamily-specific domain, such as hATC (*hAT* C-terminal dimerization) for *hAT* and SWIM (SWI2/SNF2 and *MuDR*) for MULE. For some TIR-TE superfamilies, like Harbinger and CACTA, the putative DBD is encoded by another ORF distinct from the TPase-encoding ORF. Adapted from Feschotte (2008), Lizamore (2013), Schulman (2013) and Wicker et al. (2007).

### 1.2.3 Autonomous and non-autonomous TEs

Both type I TEs and type II TEs have autonomous and non-autonomous elements. An autonomous TE is capable of self-sufficient transposition by encoding all the enzymes necessary for its own mobility, whereas the transposition of a non-autonomous TE requires the enzyme produced from autonomous TEs (Slotkin and Martienssen, 2007). For example, while autonomous retrotransposons are able to generate enzymes for self-transposition, the most common non-autonomous retro-elements in plants—SINEs—require enzymes encoded by other autonomous LINE retrotransposons to accomplish mobilization (Lisch, 2013).

Similarly, autonomous type II elements, such as *CACTA*, *hAT*, and MULE superfamilies in plant genomes, comprise sequences encoding transposase surrounded by TIRs. Yet once these TEs lose their ability to produce a functional transposase due to mutations, they become the non-autonomous versions of DNA transposons. As a result, non-autonomous type II elements, composed of a pair of TIRs flanking non-transposon sequences (Slotkin and Martienssen, 2007), are usually considered as deletion derivatives of autonomous TEs or DNA elements sharing similarity with autonomous TEs only at their termini (Lisch, 2013). Mobilization of non-autonomous type II elements requires transposase produced from the autonomous elements of the same families (Slotkin and Martienssen, 2007). Miniature inverted-repeat transposable elements (MITEs) are the most common non-autonomous DNA transposons in plants (Lisch, 2013), which are characterized with only two TIRs connected in a tail-to-tail arrangement.

## 1.3 Epigenetic control of TE activity

Transposable elements are present at varying levels in different eukaryotic genomes. In plants, the prevalence of TEs ranges from 3% in the 82MB *Utricularia gibba* genome (Ibarra-Laclette et al., 2013) to 85% in the 2.3 Gb *Zea mays* genome (Schnable et al., 2009) and 90% in the 42 Gb *Fritillaria* Lily genome (Ambrožová et al., 2011; Metcalfe and Casane, 2013), whereas *Vitis vinifera* and *Arabidopsis thaliana* shows about 40% and 15% of TEs in their genomes respectively (Fultz et al., 2015; Jaillon et al., 2007). Although there are abundant of TEs in plant genomes, their transcription and transposition activity are controlled and tuned by an exquisite silencing system evolved to neutralize the mutagenic potential of TEs for host cells (Fultz et al., 2015).

In plants, there are two major routes responsible for the onset of TE silencing based on the presence or absence of small interfering RNAs (siRNAs) derived from silenced TEs, while repressive modifications on histones function to reinforce and maintain TE silencing by facilitating the compact chromatin structure that prevents TE's access to transcription machinery (Fultz et al., 2015; Sigman and Slotkin, 2016).

### 1.3.1 Canonical RNA-directed DNA methylation

Plant-specific RNA polymerase IV (Pol IV) and V (Pol V) play significant roles in driving and maintaining the upstream and downstream parts of RNA-directed DNA methylation (RdDM), respectively, via a mechanism based on the homology between the silenced TE loci and target TE sequences (Fultz et al., 2015). Pol IV can be recruited to heterochromatic loci (e.g. silenced TE loci) that have been marked by methylated cytosine and di- or tri-methylated histone H3 at lysine 9 (Law et al., 2013). RNA Dependent RNA Polymerase 2 (RDR2) then uses the transcripts produced by Pol IV as a template to generate double-stranded RNAs (dsRNAs), which are further digested by Dicer-Like 3 (DCL3) into 24 nt siRNAs (Borges and Martienssen, 2015; Cuerda-Gil and Slotkin, 2016; Fultz et al., 2015). In the latter part of the RdDM cycle, 24-nt siRNAs are captured by Argonaute 4 (AGO4) or AGO6 that collectively guide the complex to dock at the nascent RNAs transcribed by Pol V from other heterochromatic loci that share homologous sequences with the source of 24 nt siRNAs (Borges and Martienssen, 2015; Cuerda-Gil and Slotkin, 2016). The interaction of AGO4 or AGO6 and nascent RNAs attracts histone and DNA methyltransferases, mainly Rearranged Methyltransferase 2 (DRM2), to the targeted heterochromatic loci for the maintenance or re-enhancement of silencing marks, H3K9me and methyl cytosine (Borges and Martienssen, 2015; Cuerda-Gil and Slotkin, 2016).

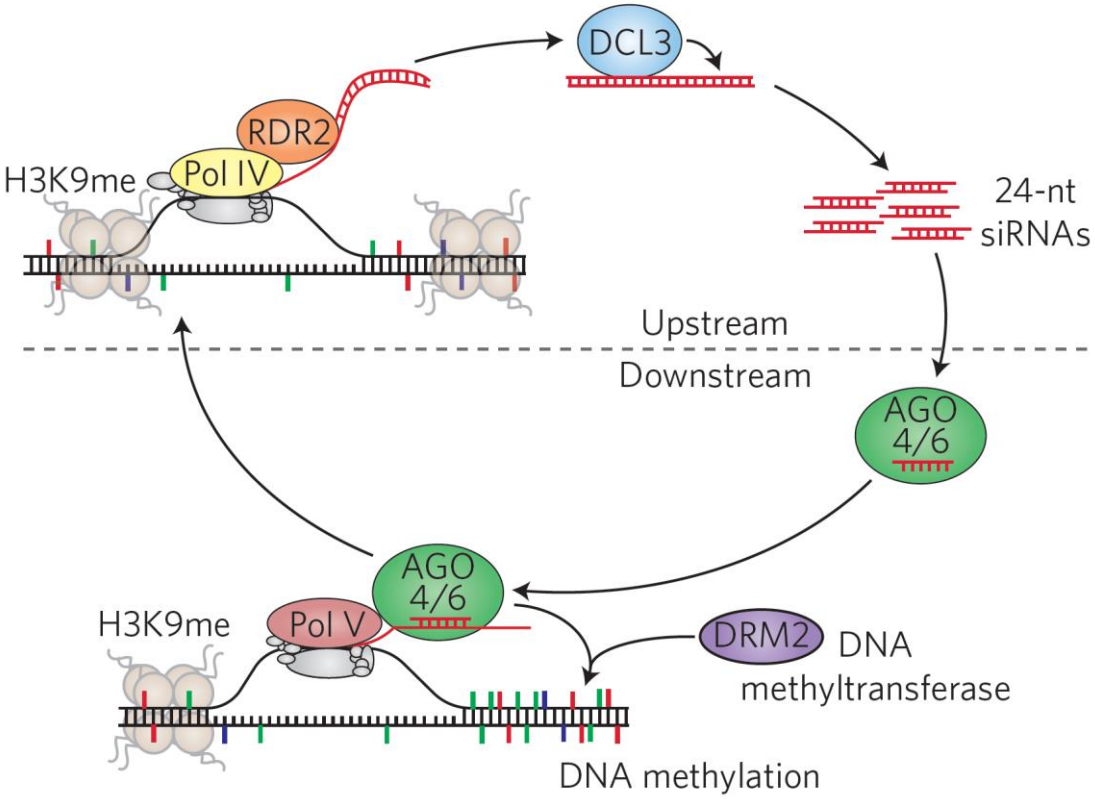



**Figure 1.2 Canonical RdDM pathway in plants**

In the RdDM pathway, heterochromatic region (double black strands) labelled with H3K9me2 and methylcytosine is bound by Pol IV and RDR2 complex to produce dsRNA (red strands), which is trimmed by DCL3 into 24 nt siRNAs. These siRNAs direct AGO4 or AGO6 to the Pol V transcript (nascent RNA) of another heterochromatic locus based on the sequence homology between the 24 nt siRNA and target loci. DNA methyltransferase DRM2 is subsequently recruited to the target loci, resulting in CHH methylation. See section 1.3.1 for more information. Reprinted by permission from Springer Nature Customer Service Centre GmbH: Springer Nature, Nature Plants, Non-canonical RNA-directed DNA methylation. Cuerda-Gil, D. and Slotkin, R.K., copyright (2016).

### 1.3.2 Post-transcriptional gene silencing

As the first layer of defence against active TEs in host cells, TE-derived 21 to 22-nt siRNAs as well as RNA Dependent RNA Polymerase (RDR6), Dicer-Like 2, DCL4 and Argonaute 1 (AGO1) constitute the RNAi system for post-translational gene silencing (PTGS; Fultz et al., 2015; Marí-Ordóñez et al., 2013). Transcriptional activation of TEs can lead to the accumulation of the primary 21-22 nt siRNAs derived from active TEs. AGO1 further incorporates these siRNAs to capture TE mRNAs, potentially leading to the cleavage of mRNAs (Figure 1.3). RDR6 then utilizes cleaved TE transcripts to generate dsRNAs, while DCL2 and DCL4 degrade the dsRNAs into secondary 21-22 nt siRNAs, resupplying resources for the next silencing cycle and reinforcing the amplification of RNAi *cis*-silencing loop (Figure 1.3; Dowen et al., 2012; Slotkin et al., 2009; Yu et al., 2013).

For the *de novo* deposition or re-establishment of silencing hallmarks on the transcriptionally active TEs, the PTGS can further trigger the canonical RdDM pathway targeting the active TE loci. Once the RNAi loop and PTGS have been established against active TEs, the dsRNA produced by RDR6 can be processed into either 24 nt or 21-22 nt siRNAs depending on the presence of DCL3 or DCL4/DCL2, respectively. (Borges and Martienssen, 2015; Cuerda-Gil and Slotkin, 2016) It is proposed that these siRNAs incorporated with AGO4/AGO6 can further recognize Pol IV/Pol V transcribed-nascent RNAs of the active TEs and facilitate *de novo* DNA methylation by recruiting DNA methyltransferase DRM2, conferring silencing hallmarks on active TEs for further establishment of chromatin modification (Fultz et al., 2015; Marí-Ordóñez et al., 2013; McCue et al., 2015). Depending on the homology between the transcriptionally active TE locus and its inactive counterparts (e.g. silenced TE loci of the same family), it is likely that the RdDM initiated by PTGS that defends the active TE locus might also target its silenced TE counterparts, thus enhancing the epigenetic suppression on these sites.

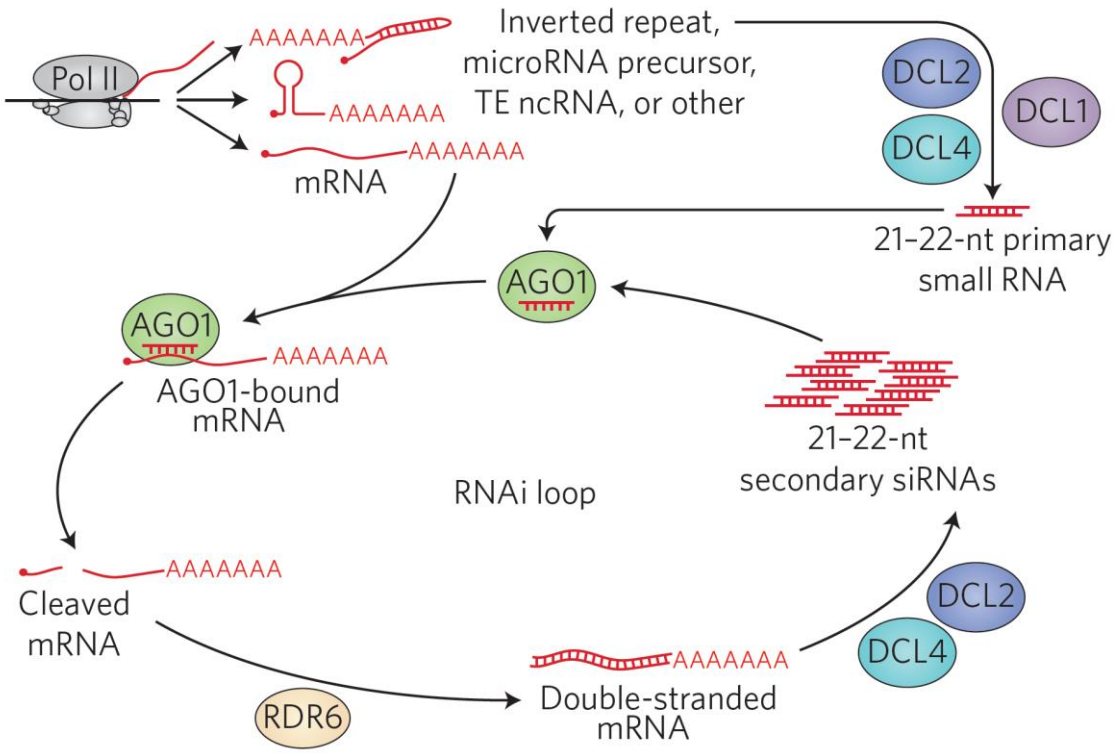



**Figure 1.3 PTGS in plants**

Based on the sequence context, Pol II transcripts produced from inverted repeat, microRNA or TE loci can self-fold into dsRNA, which is further processed into 21-22 nt primary siRNA by DCL1, DCL2 and DCL4. These primary siRNAs guide AGO1 to cleave the mRNA product derived from the active loci. The cleaved mRNA serves as the template for RDR6 to generate dsRNA, which is further trimmed by DCL2 and DCL4 into 21-22 nt secondary siRNAs that can be subsequently fed into the self-magnified RNAi loop to target additional copies of the mRNA for cleavage and amplify the production of secondary siRNAs. See section 1.3.2 for more details. Reprinted by permission from Springer Nature Customer Service Centre GmbH: Springer Nature, Nature Plants, Non-canonical RNA-directed DNA methylation. Cuerda-Gil, D. and Slotkin, R.K., copyright (2016).

### 1.3.3 Chromatin modification

Most silenced TEs are tightly wrapped around histones having specific modifications in amino (N)-terminal tails, such as dimethylated histone H3 at lysine 9 (H3K9me2), monomethylated histone H3 at lysine 27 (H3K27me1) and monomethylated histone H4 at lysine 20 (H4K20me1), that collectively compact associated DNA into heterochromatin (Roudier et al., 2011). While these histone modifications are often associated with TE DNA, H3K4me and H3K36me are frequently appear in promoters or gene bodies of active genes, thus maintaining the euchromatic chromatin structure (Zentner and Henikoff, 2013). In addition, histone acetylation has also been reported as a key modification indicative of permissive transcription of DNA associated with this histone mark. Unlike histone methylation marks, the accumulation of acetyl groups on histone tails is positively correlated with the transcriptional activity of underlying genes (Shahbazian and Grunstein, 2007; Zentner and Henikoff, 2013). Promoter regions of active or inducible genes are characterized by enrichment of acetylated histone. They are preferentially bond by histone acetyltransferase (HAT) and histone deacetylase (HDAC), suggestive of the histone acetylation/deacetylation circuit in favour of the accumulation of acetylated histone marks and, therefore, the ability for the underlying DNA to become transcriptionally active (McAnena et al., 2017). In contrast, heterochromatic regions are often inadequate in acetylated histone marks (Kurdistani and Grunstein, 2003).

The histone marks of heterochromatin can be established following the aforementioned Pol IV-RdDM pathway. The H3K9me mark laid by RdDM can guide Methyltransferase 1 (MET1) to replicate CG DNA methylation during cell division (Mathieu et al., 2007), whereas Chromomethylase 2 (CMT2) and CMT3 act to sustain heterochromatic CHH and CHG methylation (where H=A, T or C) respectively (Stroud et al., 2014). Methylation of TE sequences leads to further deposition of H3K9me2 by histone methyltransferase KYP/SUVH4 (Kryptonite/Su(var)3-9 homolog 4), SUVH5 and SUVH6 via recognition of methyl CHG or CHH (Ebbs and Bender, 2006; Jackson et al., 2004; Johnson et al., 2007), while SUVH2 confers H3K9me2, H4K20me1 and H3K27me2 modifications via interactions with methylated CG sites (Naumann et al., 2005). In addition to pathways mediating suppressive histone methylation, it has been reported that *Arabidopsis* HDAC AtHAT6 can participate in TE silencing by interacting with MET1 and SUVH4/5/6 (To et al., 2011a; Yu et al., 2017). Given the function of SUVH4/5/6 and HDAC in maintaining suppressive chromatin structure, it is sensible to speculate that pharmacological inhibition of these enzymes might interfere with the homeostasis of chromatin conformation, leading to a TE transcriptional storm.

**Table 1.1 Summary of DNA methylation enzyme in plants**

| Protein | Methylation context | Preferential involving pathway | Reference |
|---|---|---|---|
| DRM2 | CHG<br>CHH<br>(CHH preferentially) | *De novo* non-CG methylation in RdDM | (Stroud et al., 2014; Zemach et al., 2013) |
| MET1 | CG | Maintaining CG methylation pattern during DNA replication | (Deniz et al., 2019; Mathieu et al., 2007; Zemach et al., 2013) |
| | | CG methylation of DDM1-mediated heterochromatin silencing | |
| CMT2 | CHG<br>CHH<br>(CHH preferentially) | Non-CG methylation of DDM1-mediated heterochromatin silencing | (Stroud et al., 2014; Zemach et al., 2013) |
| CMT3 | CHG<br>CHH<br>(CHG preferentially) | Maintaining CHG methylation pattern during DNA replication | (Lindroth et al., 2001; Stroud et al., 2014; Zemach et al., 2013; Zhang et al., 2018) |
| | | Non-CG methylation of DDM1-mediated heterochromatin silencing | |

**Table 1.2 Summary of histone methyltransferase maintaining heterochromatic marks in plants**

| Protein | Methylation context | Interacting cytosine methylation | Reference |
|---|---|---|---|
| SUVH2 | H3K9me2<br>H4K20me1<br>H3K27me2 | Methylated CG | (Naumann et al., 2005) |
| SUVH4 | H3K9me2 | Methylated CHG or CHH | (Ebbs and Bender, 2006; Jackson et al., 2004; Johnson et al., 2007) |
| SUVH5 | | | |
| SUVH6 | | | |

In addition to DNA methylation and histone modification, the swi/snf chromatin-remodeling protein DDM1 plays a crucial role in maintaining the DNA methylation level of heterochromatic TEs via interaction with histone H1, leading to increased accessibility of DNA methyltransferases (MET1, CMT2 and CMT3) to the heterochromatic TE loci for the catalysis of cytosine methylation (Zemach et al., 2013). DDM1 functions as a master regulator required for methylation of DNA and H3K9 in silenced genes and TEs. Arabidopsis ddm1 mutant not only shows a dramatic decrease of DNA methylation but also reveals a replacement of methylated H3K9 by methylated H3K4, a permissive mark for transcriptional activation of underlying DNA (Gendrel, 2002). These changes also underpin the increased TE mobilization observed in *ddm1* mutant (Hirochika et al., 2000; Miura et al., 2001; Singer, 2001), demonstrating that maintenance of condensed chromatin state via DNA and histone methylation is pivotal to TE silencing.

## 1.4 Roles of TEs within genes

The majority of TE-rich heterochromatic regions contribute to the centromere, pericentromeres and TE islands (Sigman and Slotkin, 2016). Nevertheless, a substantial number of TEs locate within genes, primarily at introns (Saze et al., 2013). Saze and colleagues (2013) found several thousand rice genes possessing long introns with heterochromatin features suggestive of the presence of TEs within these introns. A study from West et al. (2014) revealed that more than 10% of maize genes have at least one TE insertion within an intron. In *Arabidopsis*, about 3% of TEs reside within gene units, whereas 85% of them are intronic TEs (Le et al., 2015). Moreover, a recent study of an LTR retrotransposon *LORE1* activated during Lotus japonicus germline propagation observed 68% of insertions were intragenic (Małolepszy et al., 2016). As opposed to most intergenic TEs that are often concealed in numbers of repetitive sequence islands, intragenic TEs are more accessible to study the mechanisms regulating TE transcriptional activity and mobility as well as the interaction between such TEs and the host genes.

As most exonic TE insertions usually result in disruption of gene function, this introduction has focused on TE insertions within introns and promoters.

### 1.4.1 TEs within introns

While TE insertions within exons frequently lead to deleterious mutations and thus selected against within a population, TEs inserted within the intronic sequence are generally tolerated by the host genome if the function of the affected host gene is not interrupted (Sigman and Slotkin, 2016). With low selection pressure, intronic TE sequences may potentially decay over time or undergo excision and deletion (Sigman and Slotkin, 2016). By way of example, over 80% of intragenic TEs in *Arabidopsis* are shorter than 1kb, while at the same time, greater than 50% of intergenic TEs are longer than 1kb (Le et al., 2015). It implies that intragenic TEs are relics that are truncated or degenerated from ancestral TE sequences (Le et al., 2015). In general, intragenic *Arabidopsis* TEs are less methylated than intergenic TEs yet are more methylated than other genic regions (Le et al., 2015). However, the level of non-CG methylation is similar between intragenic and intergenic TEs in maize (West et al., 2014), indicating wide variation between species, genus and families.

In maize, rice, and *Arabidopsis* genomes, most intragenic TEs are within introns with the appearance of heterochromatic hallmarks (Saze et al., 2013; West et al., 2014; Le et al., 2015). Nevertheless, controversy exists over the effect of intronic TEs on host genes (Hirsch and Springer, 2017). In *Zea mays*, intragenic TEs with heterochromatic marks do not restrict the expression of host genes, as some of these TEs are within introns (West et al., 2014). In Norway spruce, the expression of genes is not influenced by intronic TE insertions, although these TEs seem to be transcriptionally inactive

(Nystedt et al., 2013). In contrast, *Arabidopsis* genes possessing intronic TEs are transcribed at a lower level than genes without TEs, and the expression level of host genes is negatively associated with the methyl CHG level of the intronic TEs (Le et al., 2015). Indeed, DNA methylation or suppressive H3K9me2 of intronic TEs are responsible for the use of correct transcription termination sites and splicing sites of *Arabidopsis* host genes (Lei et al., 2014; Sigman and Slotkin, 2016; Tsuchiya and Eulgem, 2013), while epigenetic mutants, such as *cmt3* and *ddm1*, showed a significant association between reduction of non-CG methylation at intronic TEs and the increase of premature transcription termination of host genes (Le et al., 2015).

The balance between suppressive histone marks H3K9me2 of intronic TEs and permissive H3K4me3 of protein-coding regions may illuminate the link between intronic TE activity and host gene expression levels. The H3K9 demethylase IBM1 (Increase in *BONSAI* Methylation 1) is responsible for containing H3K9me2 within TE regions by pruning off the methyl groups of H3K9me2 that spread into gene coding regions (Saze et al., 2008). In contrast, the histone demethylases JMJ14 (Jumonji 14), LDL1 (Lysine-Specific Demethylase 1-Like 1), and LDL2 act to prevent the spread of H3K4 methylation into heterochromatic TEs by removing the methyl groups at H3K4 from histones wrapped by TE sequences (Greenberg et al., 2013). In addition to the H3K4 and H3K9 boundary maintenance mechanism, DNA glycosylase proteins ROS1, DME, DML1, and DML2 function to enzymatically remove 5-methylcytosine in transcriptionally permissive areas, and therefore prevent the spread of DNA methylation into active gene regions (Penterman et al., 2007; Zhu et al., 2007).

Several researchers have demonstrated that a loss of epigenetic silencing marks on intronic TEs is responsible for aberrant alternative splicing and polyadenylation. In oil palm, the “mantled” abnormality of the fruit, a somaclonal variation arising from tissue culture, leads to reduced fruit yield and oil production. Ong-Abdullah et al. (2015) discovered that this phenotype results from hypomethylation of the *Karma* retrotransposon in the fifth intron of the homeotic gene *DEFICIENS* (Ong-Abdullah et al., 2015). The hypomethylation phenotype is associated with the loss of 24-nt siRNAs and an increase of a truncated transcript of the *DEFICIENS* gene due to a splicing aberration extending from the donor site within the 5$^{th}$ intron to the proximal *Karma* acceptor site. In *Arabidopsis*, the histone-binding protein EDM2 (Enhanced Downy Mildew 2) binds with heterochromatic introns and recognizes both the suppressive H3K9me and the permissive H3K4me marks (Lei et al., 2014). It has been found that, for genes containing heterochromatic TEs in introns, EDM2 is essential for the production of functional mRNA transcripts of these genes by enhancing splicing of intronic TEs, transcriptional elongation, and use of distal polyadenylation sites (Deremetz et al., 2019; Saze, 2018). Loss of EDM2 leads to a reduction of H3K9me2 levels at a retrotransposon, *COPIA-R7*, that is found within the first intron of the *Arabidopsis* disease resistance gene *RPP7*, drives

the accumulation of a non-*RPP7*-coding isoform using the alternative polyadenylation site at the 5' LTR of *COPIA-R7* (Tsuchiya and Eulgem, 2013).

Taken together, heterochromatic hallmarks laid on intronic TEs are crucial for the repression of TE transcriptional activity and the prevention of aberrant transcription of the host genes. These genes, however, are less likely to be highly expressed than genes without TEs, possibly due to the suppressive epigenetic marks in their introns. This assumption leads us to the speculation that the host cell might balance the level of epigenetic suppression against intronic TEs and the transcriptional activity of host genes; while epigenetically unmasked intronic TEs are potentially mutagenic and might interfere with proper transcription of the host genes, intense epigenetic repression of these TEs is likely to limit the expression of the related genes. Nonetheless, as discussed below, TEs at promoters of host genes might impact the expression of these genes in more diverse ways than intronic TEs.

### 1.4.2 TEs at promoters

A study in maize utilized the cap analysis gene expression (CAGE) technique to map transcription start sites (TSSs) for over 17,000 genes and found that 180 of these annotated TSSs in the maize genome overlap with TE-derived sequences (Hirsch and Springer, 2017; Mejía-Guerra et al., 2015), demonstrating a substantial prevalence of TEs within, or proximal to, promoter regions in host genomes.

Despite the dependency on the host cell's transcriptional machinery (e.g. RNA Pol II), most autonomous TEs have their own promoters to initiate their transcription. To gain the opportunity for transcriptional activation, these TEs have evolved *cis*-regulatory sequences that can be recognized by the host cell's transcription factors activated upon various developmental and differentiation signals and external stimuli (Chuong et al., 2017). TEs integrating at promoter regions of genes can provide novel *cis*-elements and lead to exaptation of stress-responsiveness for host genes. Using LTR retrotransposons as an example, the TSS and *cis*-regulatory element harboured within the identical LTR promoter regions can drive both the expression of the TE itself and the downstream DNA sequences from the 5' and 3' LTRs, respectively.

In *Citrus sinensis*, the *Ruby* gene is a transcription factor responsible for anthocyanin biosynthesis, which is rarely expressed in the blond orange variety Navalian (Butelli et al., 2012). In wild type plants, anthocyanin biosynthesis is usually limited to the fruit flesh. In the blood orange variety, Tarcco, a *Copia*-like retrotransposon, *Rider*, was found inserted immediately upstream of the *Ruby* coding sequence, providing new *cis*-regulatory elements that drive an alternate transcriptional programme resulting in the production of the distinctive red flesh of this blood orange variety (Butelli

et al., 2012; Lisch, 2013). The TE-derived element not only impacts *Ruby*'s spatial transcription but also contributes to *Ruby*'s cold-responsiveness. Furthermore, recombination between the LTRs leaves a solo LTR that further enhances *Ruby* expression in the variety Maro (I) (Butelli et al., 2012).

In an Indonesian rice cultivar Tjahaja, the 3'LTR of the retrotransposon *Renovator* inserted at the promoter region of the rice blast resistance gene, *Pit.* The *Renovator* element contributes novel promoter regulatory elements and contributes to *Pit* reactivation in response to the fungus *Magnaporthe grisea*, conferring the pathogen responsiveness of *Pit* and disease resistance of the cultivar (Hayashi and Yoshida, 2009). Conversely, the *Pit* gene in the susceptible rice cultivar, Nipponbare, lacks this TE insertion upstream of *Pit* using its original promoter to drive transcription (Hayashi and Yoshida, 2009). The transcription level of Tjahaja's *Pit* gene was 34-fold higher than that of Nipponbare's *Pit* gene upon pathogen inoculation (Hayashi and Yoshida, 2009). Comparison of the *Pit* promoter activity between these two cultivars using *GUS* reporter gene assay reveals that the novel promoter comprised of *Renovator* 3' LTR contains sequence functioning as a transcription enhancer and exhibits significantly higher activity than *Pit*'s own promoter (Hayashi and Yoshida, 2009).

Similarly, the insertion of a type II transposon *Mutator* within the promoter of the maize gene *hcf106* can provide read-through transcription of *hcf106* from cryptic TSSs of the transposons. Interestingly, the cryptic promoter is only effective when the TE is silenced, whereas the active TE disrupts *hcf106* expression (Barkan and Martienssen, 1991; Hirsch and Springer, 2017; Settles et al., 2001). As opposed to the case in maize, Yang et al. (2005) highlighted a multifaceted role for rice TEs within promoters with respect to host gene expression in terms of the transcriptional enhancement of TEs versus the suppressive DNA methylation state of TEs. In rice (*Oryza sativa*) line IR24, the promoter of *ubiquitin2* (*rubq2*) gene contains a DNA transposon *Kiddo* nested within another transposon *MDM1*, whereas the *urbq2* promoter in the subspecies *japonica* line T309 is essentially identical to that in IR24 except for the absence of *Kiddo* insertion (Yang et al., 2005). The *urbq2* gene is expressed at a similar level in these two rice lines. However, promoter activity assay utilizing the *GFP* reporter gene reveals that the GFP signal driven by the promoter containing *Kiddo* is 1.25 times higher than that driven by the promoter lack of *Kiddo* insertion (Yang et al., 2005). Although this *Kiddo* insertion can elevate the promoter activity, it is endogenously targeted by DNA methylation that has likely neutralized its enhancement to the transcriptional activation of *urbq2*, since hypomethylation of the TEs, generated via 5-azacytidine treatment, resulting in a threefold increase in *rubq2* transcription in line IR24 but not in line T309 (Yang et al., 2005).

These observations suggest that TEs at promoter regions are often targets for epigenetic modification. However, the impact of TEs within or near promoters is variable and should be assessed on a case-by-case basis.

## 1.5 Activation of TEs during stress response in plants

Despite the fact that TE transcription is predominantly suppressed by epigenetic silencing involving DNA methylation, histone modifications and siRNAs, TE transposition can be triggered by exposure to biotic and abiotic elicitors (Lisch, 2013). The consequences are not only the accumulation of new TE insertions but also changes of the epigenetic landscape shaped by the transcriptional activation of the pre-existing TE loci and the distribution of new insertions. Due to their multifaceted impact on genomes, TEs are a very attractive mutagen, aside from being an endogenous agent for insertional knock-outs. The apparent increase in TE activity under a given set of conditions is important in that it may offer a means to heighten TE activity and, therefore, mobilisation, which is the ultimate goal of the thesis.

A DNA transposon, *Tam3,* in *Antirrhinum majus* rarely transposes at 25˚C, but these elements are more frequently mobilised at 15˚C. Hypermethylation of *Tam3* has been observed in plants grown at higher temperatures. As temperatures are reduced, there is a reduction of methylation at *Tam3* loci associated with *Tam3* activation (Hashida et al., 2006). Studies in tobacco (*Nicotiana tabacum*) found that the LTR retrotransposon *Tnt1* contains three subfamilies (*Tnt1A*, *Tnt1B* and *Tnt1C*) that are diverse in their U3 sequence within LTR (see the LTR structure in Figure 1.1 A) and respond to different elicitors. While *Tnt1A* subfamily is strongly induced by the fungal elicitor cryptogein and methyl jasmonate (MJ), a defence-related signalling molecule induced by wound and infection by necrotrophic pathogens, *Tnt1C* is more sensitive to the phytohormone salicylic acid (SA) that can be induced by biotrophic pathogens (Beguiristain et al., 2001). A *copia* type LTR retrotransposon *ONSEN* in *A. thaliana* is heat-sensitive due to a number of heat-responsive elements found within its promoter, which can be recognized by heat stress defence factors derived from hosts, resulting in an increase of *ONSEN* RNA level and the accumulation of its extrachromosomal dsDNAs that are likely to be the reverse transcription products ready for integration (Cavrak et al., 2014). Another *Arabidopsis* TE that partially overlaps with the 5' UTR of the gene locus *At4g39860* was found transcriptionally activated in the presence of SA (Dowen et al., 2012). The activation of this TE was accompanied by a reduction of DNA methylation in the TE sequence and an accumulation of 21 nt siRNAs derived from it. The transcription level of the co-localized *At4g39860* was co-activated with this TE as well (Dowen et al., 2012).

In addition to biotic and abiotic stimuli, plant embryogenic cell cultures have also been shown to associate with reactivation of transposon activity, in terms of both transcriptional activation and mobilization, in correlation with epigenetic changes. For example, in *Medicago truncatula*, the transcriptional activity of the Copia-type retrotransposon *MERE1* is heightened during tissue culture,

so is its mobility indicated by increased insertion copies (Rakocevic et al., 2009). It is correlated with a decrease of methylation across these retroelements (Rakocevic et al., 2009).

Taken together, these observations support that TEs can be activated under the control of self-enclosing *cis*-regulatory elements, which act to impart abiotic, biotic and tissue culture transcriptional regulation of these elements (Grandbastien, 1998; Takeda et al., 1999). Moreover, some elements also exhibit a clear correlation between observed TE transcriptional activation and altered epigenetic marks at TE loci, including a reduction in DNA methylation and accumulation of siRNAs mediating PTGS. It suggests that relaxed epigenetic regulation of TE loci leads to increasing the potential for these loci to become transcriptionally active. By investigating transgenerational activation of TE transcription, Marí-Ordónez et al. (2013) further demonstrated the association of TE activation with dynamic changes of DNA methylation and siRNA level in *Arabidopsis*. At the early stage of retrotransposon *Evadé* (*EVD*) activation, high expression of TEs is accompanied by low DNA methylation and 24 nt siRNA levels, regardless of the presence of 21-nt siRNAs. While *EVD* copy number increased as a consequence of *EVD* activation, the expression level of *EVD* declined in association with accumulation of 24-nt siRNAs and establishment of DNA methylation, featuring transcriptional gene silencing led by RdDM.

## 1.6 Hypotheses

### 1.6.1 $H_1$: It is possible to distinguish a subset of transcriptionally active TE loci.

Competent *de novo* transcription of autonomous TEs is of great importance for TE mobilization irrespective of TE types. The self-proliferating nature of TEs presents several complicating issues for the identification and study of these loci. TE-derived sequencing reads from short-read sequencing technologies (e.g. 150bp pair-end read technology from Illumina) can map to multiple positions in the reference genome and confound the alignment results. Consequently, investigation of transcriptionally active TEs is often restricted at the family level, while identifying individual transcriptionally active loci is considered difficult or impossible (Jin et al., 2015; Shahid and Slotkin, 2020).

At the outset of this research, long-read or ultra-long read sequencing technology was not available or affordable. Even with the assistance of long-read sequencing technology, due to high levels of sequencing error and the self-propagating nature of TEs, TE transcripts derived from highly conserved yet truncated TE loci might still overlap with other identical or highly similar loci when analysing the data. Because of their mutagenic properties and the abilities to modulate gene transcription, as previously mentioned in section 1.4, TEs are frequently silenced by epigenetic mechanisms. This characteristic has facilitated the accumulation of nucleotide substitution or small insertions and deletions in the silenced TEs. In contrast, TEs from more recently active TE family would have accumulated fewer variations and thus more identical than the individual loci of more ancient TE family. The mutational erosion of TE sequences preserves the lineage information and transposition history of TEs, providing leverage to identify subsets of transcriptionally active TE loci in the genome. Therefore, with a similarity-based approach, it is plausible to identify TE families with the most recent mobilization burst, capture TEs with transcriptional potential from the silenced loci, and sort these potentially active loci by their sequence variances.

*$H_{0\text{-}1}$: It is impossible to tell the difference between the transcriptionally active TE loci and the silenced ones.*

### 1.6.2 $H_2$: The position of TEs within genes can reveal the transcriptional activity of TEs

Intragenic TE insertions have been observed in exons, introns, promoters, and untranslated regions (UTRs) of host genes (Sigman and Slotkin, 2016), and the various phenotypic effects on host cells are widely discussed. In general, exon insertions usually result in disruption of coding sequences representing detrimental mutations eliminated through natural or purifying selection (Sigman and Slotkin, 2016). However, host cells are more likely to tolerate intronic TE insertions (Hirsch and

Springer, 2016), as long as the insertions do not cause severe disturbance to host gene expression through mechanisms such as alternative splicing and premature polyadenylation. TE insertion within promoter regions can contribute to the acquisition of novel promoters and the exaptation of biotic or abiotic responsiveness for the host genes (Lisch, 2013). Both intronic and promoter TEs have been found constrained by epigenetic silencing, yet the silencing level and the silencing effect to TEs and host genes can be different for different insertions and insertional contexts. Intriguingly, although TEs at UTRs are less documented in plants, Kabelitz and coworkers (2014) observed that intragenic TEs at 3'UTRs were highly expressed in cells as the degree of epigenetic silencing were low in the loci. As TEs inserted into a different part of host genes may affect the gene function in various degrees, and thus provoking different levels of silencing control to minimize or neutralize the unfavourable effects, it is possible that the intragenic TE position is associated with TE activity.

The null hypothesis is:

*$H_{0\text{-}2}$: The position of TEs within genes is not associated with the transcriptional activity of TEs.*

### 1.6.3 $H_3$: The transcriptional activity of TEs co-localized with genes is associated with the activity of the corresponding host genes.

Although most TEs are tightly packed into gene-poor heterochromatin, a substantial number of genic TEs insertions have been discovered in both monocotyledonous and dicotyledonous plant species, such as rice (Saze et al., 2013), maize (West et al., 2014), *Arabidopsis thaliana* (Le et al., 2015), and *Lotus japonicus* (Małolepszy et al., 2016). These studies raise the questions whether, and how, the expression of intragenic TEs is associated with the host genes. Deep silencing of intragenic TEs may cost cells important host genes co-repressed with the TEs. To retain basal function, host cells may tune the silencing level acting on intragenic TEs, permitting TE expression.

The null hypothesis is:

*$H_{0\text{-}3}$: No significant relationship exists between the transcriptional activity of TEs and corresponding host genes.*

### 1.6.4 $H_4$: Inhibition of HDACs, which are keys to maintaining compact chromatin structure, can facilitate TE re-activation.

The condensed chromatin structure functions as a stringent physical inhibitor of TE transcription via restriction of access of the transcription machinery to TE transcription initiation sites. For instance,

*Arabidopsis* deficient in the chromatin re-modeller *DDM1* failed to maintain heterochromatin structure and suppress TE mobilization (Lee et al., 2020; Zemach et al., 2013).

The chromatin structure is modulated chiefly by modifying histone tails, which is a dynamic balance between transcriptionally permissive and restrictive modifications. Section 1.3.3 has discussed the common types of histone modification related to TE activity. Histone methylation at different lysines of the histone tail can result in distinct transcriptional regulation (e.g. H3K4me is permissive, whereas H3K9me is suppressive). In contrast, histone acetylation is often associated with transcriptionally permissive region irrespective of the position of acetylated lysine on the N-terminal tail of H3 and H4 (Zentner and Henikoff, 2013).

Studies in yeast and mouse cell lines reveal that acetylated histone on promoter and gene body serves as signature to be recognized by RNA Pol II, thus facilitating transcription initiation and elongation (Stasevich et al., 2014; Wang et al., 2009). By contrast, acetylated histones are rarely found in heterochromatin (Wang et al., 2009; Zentner and Henikoff, 2013). Genome-wide reprograming of histone acetylation has been observed through the establishment of plant tissue culture (Law and Suttle, 2005; Tanurdzic et al., 2008), temperature shift during seedling development (Hu et al., 2012; Tittel-Elmer et al., 2010), and wound treatment of *Arabidopsis* roots (Rymen et al., 2019). Subsets of transcriptionally reactivated genes and TEs identified in these studies are linked to the hyper-acetylation of histones associated with these regions, implicating an opposing role of histone acetylation on transcriptional silencing.

It has been proposed that, during DNA replication, acetylated H3 and H4 are non-specifically deposited onto newly synthesized genomic DNA and collectively wrapped into nucleosome along with the histone dimers H2A/H2B (Shahbazian and Grunstein, 2007). Soon after nucleosome assembly, histone deacetylase (HDAC) selectively remove the acetyl groups from histones laid on specific genomic regions, thus facilitating the establishment of heterochromatin (Shahbazian and Grunstein, 2007). This model hence suggests that impairment of HDAC function might halt the formation or maintenance of heterochromatin structure.

Altogether, these data give rise to the idea that the blockage of the dynamic histone de-acetylation circuit might de-repress TE activity. Pharmacological inhibitors of histone deacetylase (HDACi) have been widely tested in patients, animal models and plants (Bolden et al., 2006; Falkenberg and Johnstone, 2014; Ma et al., 2013). Relaxed chromatin structure has been observed in organisms treated with HDACi, while *Arabidopsis* mutant *hda6* show increased transcriptional activity of TEs (Liu et al., 2012; Yu et al., 2017). Therefore, drug inhibition of HDACs might result in TE perturbation.

*$H_{0\text{-}4}$: TE transcriptional activity cannot be stimulated by inhibition of HDACs.*

### 1.6.5 $H_5$: TE perturbation due to HDACi can, in turn, enhance PTGS or RdDM.

The epigenetic systems in plants comprise multiple interwoven pathways to suppress TE transcriptional activity or intercept and destroy TE transcripts. In the *Arabidopsis* lack of functional chromatin remodeller DDM1, biogenesis of 21-22 nt siRNA dependent on RDR6 was turned up to stimulate PTGS against activated TEs (Lee et al., 2020). Likewise, deficiency in the Pol IV subunit involved in RdDM enabled the heat-responsive ONSEN retrotransposition accompanying with increased accumulation of 21 nt siRNA in heat-shocked Arabidopsis seedlings (Ito et al., 2011). In 'mantaled' oil palm, hypomethylation and loss of 24 nt siRNA on a LINE locus resulted in a spike of 21 nt siRNA (Ong-Abdullah et al., 2015). While hypomethylation was frequently observed in transcriptionally active TEs, hypermethylation of CHH was found to be re-established after the transcriptional activation (Marí-Ordóñez et al., 2013; Secco et al., 2015), suggesting re-enhanced RdDM. Although *Arabidopsis* histone deacetylase AtHDA6 was found interacting with histone methyltransferase SUVH4-6 and DNA methyltransferase MET1 (Liu et al., 2012; To et al., 2011a), HDAC's participation in siRNA biogenesis and other epigenetic pathway have been rarely reported. Overall, it is sensible to speculate that the PTGS or RdDM would not be severely impacted by pharmacological inhibition of HDACs. Instead, PTGS or RdDM silencing might be strengthened to neutralize TE transcripts or re-suppress TE transcriptional activity, respectively.

*$H_{0\text{-}5}$: HDACi-induced TE perturbation has no effect on PTGS and RdDM*

## 1.7 Summary of chapters

This thesis firstly establishes an analysis pipeline to identify potentially expressed TE loci (termed 'expression candidates') from the short-read RNA sequencing (RNAseq) data of grapevine embryogenic callus subjected to biotic stressors (chapter 2). With the ability to identify individual loci of expressed TEs, the characteristics of these TE loci (e.g. integrity, location and distinctiveness) were investigated in chapter 3, and the relationship of transcriptional activity between TEs and genes was examined in chapter 4. Therefore the results of chapter 3 and chapter 4 are discussed with regard to the prerequisite for mobilisation of TEs being their transcription and the impact that their transcription has on co-localized genes.

To validate that the pipeline is applicable beyond our grapevine model system, we applied the analysis pipeline established in chapter 2 to published RNAseq data of *A. thaliana* and *Drosophila melanogaster* (chapter 5). While our analysis pipeline (based on short-read sequencing data) improved the granularity in interpreting TE transcriptional activity to the level of individual TE loci,

long-read sequencing data is still required to obtain intact full-length information of TE transcripts. Chapter 6, therefore, utilizes long read Oxford Nanopore Technology (ONT) cDNA sequencing to validate the findings of chapter 3 and chapter 4 and investigate aberrant alternative splicing of genes associated with TEs.

It was clear from our initial data that stress treatments alone are insufficient to stimulate the mass mobilization of TEs. While mutation of components of the epigenetic silencing machinery or the use of pharmacological drugs to achieve similar ends has been shown to be effective in allowing increased TE mobilisation, the impact of directly manipulating heterochromatin structure has not been explored in this context. To test whether pharmacological inhibition of HDACs can promote TE re-activation, we exposed grapevine callus cultures with HDAC inhibitors (HDACi), trichostatin A or 4-phenylbutyric acid. The transcription patterns of TEs and genes in the presence of HDACi were investigated by Illumina and ONT cDNA sequencing (chapter 7).

Last but not least, there are unresolved questions around the epigenetic impact of altering the transcriptional profiles of TE loci. One component driver of re-silencing of transcriptionally activated TEs is the production and targeting of small RNAs that act to silence TE transcripts (through PTGS) and the concordant DNA loci via RdDM. To this end, we interrogated the accumulation changes of small RNA populations in the experiments outlined above in chapter 8.

Finally, in chapter 9, we present this study's conclusions with respect to the originally proposed hypotheses presented in this chapter (section 1.6) and conclude with work that should be carried out into the future.

# Chapter 2

# Analysis pipeline for identification of potentially expressed transposable elements

## 2.1 Overview

Numerous studies have shown that TE activity is tightly repressed by multiple epigenetic components. The epigenetic silencing system endeavours to besiege TE activity from all directions and extinguish the sparks of mobilization. Nonetheless, there is increasing evidence showing the important role of TEs in increasing genetic diversity in a population exposed to varying stressors, hence facilitating adaptation to rapidly changing environmental conditions. To harness TE biology for crop improvement, we need a more comprehensive understanding of how host cells manipulate the balance between epigenetic silencing and activation of TEs within the genome.

For this purpose, it is crucial to pinpoint loci in the genome producing TE-related transcripts, the first stage in the TE life cycle. The repetitive nature of TE sequences, however, presents multiple challenges in studying TE biology utilising second and third generation sequencing technologies. By the time we started to address the aforementioned question, the short-read sequencing platform was more affordable and widely adopted than the third generation long-read sequencing technology. A number of analysis tools, preferentially for short-read sequencing data, have been developed to capture subsets of transcriptionally active TE loci. Starting with the well-developed short read sequencing system, this chapter established a pipeline of pre-developed tools each capable of identifying particular sets of active TEs to maximize the identification of potentially active TE loci.

The application of the pipeline on short-read RNA sequencing (RNA-seq) data revealed its ability to exclude the majority of annotated TE loci based on low levels of alignment depth. This step alone narrowed the investigation from assessing hundreds of thousands of annotated TE loci to a few thousand loci. Comparison of this pipeline with the established software *TEtranscripts* showed high consistency in terms of the identification of active TE families. Utilising the presence or absence of unique-mapping reads, the potentially expressed TE loci (which we term expression candidates) were further categorized into trackable or untrackable groups, respectively. The trackable loci are likely to be older insertions having accumulated mutations that enable positive identification of transcribed loci and can therefore be used in differential expression analysis. The untrackable loci could be newly transposed loci that are relatively identical, hence contributing to the alignment of reads to multiple loci; while unsuitable for differential expression analyses, they are valuable in identifying a pool of TE

loci producing functional transcripts for autonomous mobilization. It should be noted that the goal of this pipeline was to assist in the identification of transcriptionally active TE loci rather than quantification of TE transcription.

## 2.2 Introduction

### 2.2.1 Why transposable elements matter?

Mobilization of endogenous transposable elements (TEs) is mutagenic and, therefore, usually tightly suppressed by host cells through epigenetic silencing mechanisms incorporating small RNAs, DNA methyltransferases, chromatin remodelers, and histone modification enzymes (Cuerda-Gil and Slotkin, 2016; Deniz et al., 2019; Girard and Hannon, 2008; Zentner and Henikoff, 2013). Apart from introducing new insertions that may disrupt gene function, TEs can cause misregulation of host genes by providing new cis- or trans-regulatory elements that are able to act on surrounding genetic loci. For example, acting as an alternative promoter of the fatty acid-binding protein gene *FABP7*, the transcriptional activation of a long terminal repeat (LTR) retroelement LTR2 resulted in the production of a chimeric LTR2-FABP7 transcript, which was found to associate with oncogenic progression in B cell lymphomas (Lock et al., 2014). In oil palm, demethylation of an intronic Karma TE within the *MANTLED* gene was shown to attenuate the transcription of *MANTLED*, resulting in a mutant mantled fruit phenotype (Ong-Abdullah et al., 2015). However, as 'double-edged swords', co-opted TEs have been documented as gene regulatory elements primed by transcription factors in many other cases, in which the co-localized genes acquired tissue specificity or stress responsiveness. For instance, an LTR retrotransposon was reported as an oocyte-specific promoter essential for female oocyte function and fertility in mouse (Flemr et al., 2013), and an endogenous retrovirus (ERV) element was identified as a regulatory sequence of human *Absent in Melanoma 2* (*AIM2*) gene to take part in innate immunity (Chuong et al., 2016). As an example of introducing low-temperature responsiveness to neighbouring gene, the insertion of LTR retrotransposon Rider upstream of the anthocyanin biosynthesis gene *Ruby* gave rise to the blood orange varieties cued by cold temperatures (Butelli et al., 2012). From a long-term perspective, the accumulation of TE transposition may result in an enrichment of genetic variety, thus facilitate crop improvement or adaptation to the ever-changing environment. A TE insertion event dated back to 1819 was found to promote the industrial melanism in the peppered moth (*Biston betularia*) population during the Industrial Revolution, in which the common pale moth was replaced by a previously un-reported dark-coloured form beneficial in the interaction between bird predation and air-polluted environment (Van't Hof et al., 2016). During the domestication of maize from its wild progenitor teosinte, the naturally occurred retrotransposition of Hopscotch was identified as a transcription enhancer of the domestication gene *teosinte branched1* (*tb1*), and partially contributed to the favoured apical dominance phenotype whereby development of axillary branches was suppressed and nutrients were concentrated in the main stem (Studer et al., 2011).

Growing evidence has highlighted a series of potentially important exapted roles for endogenous TEs in genomes, hence eroding the view that TEs solely exist as genomic parasites (Bourque et al., 2018). Despite recent advances in our understanding of endogenous TE biology, there are still many questions about the complex role that these elements play in gene regulation and genome evolution. The rise of second and third generation sequencing technology has revolutionised our ability to interrogate the biology and ultimately the role that TEs play in genomes of all branches of life. Indeed, the introduction of the third-generation long-read sequencing technology, coupled with reductions in expense and error rate, has contributed to great improvements in TE assemblies (Jung et al., 2019; Sedlazeck et al., 2018; Shahid and Slotkin, 2020; Wenger et al., 2019). Following this breakthrough in high-resolution identification of TEs in eukaryotic genomes, there are growing numbers of TE-oriented transcriptome analyses utilizing these long-read platforms, albeit still rare (Panda and Slotkin, 2020; Shahid and Slotkin, 2020). Because the majority of the bioinformatic tools has been developed alongside well-established short-read sequencing methods, this research started with the short-read sequencing data to explore TE activity at the locus level. This approach allowed us to compare our findings in grapevine with the existing understanding of TE biology that is largely based on similar short-read sequencing approaches. During the course of this study, the cost-effective utilization of the Oxford Nanopore (ONT) long-read system for investigation of TE transcriptional activity became available, allowing us to further explore the transcriptional activation of TEs in the grapevine. These data will be described in chapter 6.

### 2.2.2 How to identify active TEs?

#### Identification of new TE transposition

The ambiguity of how TE's activity is modulated by host cells predisposes this study to look for active TEs at a genome-wide scale. Detection of new TE insertions appearing in progeny alongside systematic surveys of polymorphic TE insertions between closely associated individuals or species can provide solid evidence of current or recent TE activation (Huang et al., 2012). Methods for detection of such polymorphisms include derivations of Allelic Fragment Length Polymorphism (AFLP) methodologies that utilise TE specific PCR primers to 'display' new polymorphic sites, and more latterly exploration of genomic re-sequencing data coupled with high throughput TE mapping analysis (O'Donnell and Burns, 2010). While identification of new TE insertions can provide robust evidence of TE activation, it is unable to reveal the origins of the transposition, thus precluding the study of mechanisms regulating activation of parent elements.

#### Identification of transcriptionally active TE loci

To pinpoint the genomic origins of transcriptionally active TE loci, a deep survey of transcriptome data, particularly polyadenylated RNA transcripts data, may help trace the source of autonomous

transposition. This applies to both type I and type II transposon families. Type I retrotransposons are dependent on reverse transcription of transcribed loci, while type II DNA transposons require transcription of element sequence encoding a functional transposase enzyme that facilitates the transposase-mediated excision of elements from competent genomic loci. Transposition of autonomous and associated non-autonomous elements cannot be achieved without transcription of a competent element within the genome. Independent of TE transposition, transcription of TE loci, containing either intact or fragmented TEs, are thought to participate in epigenetic regulation of the genome without generating new TE insertions (Bourque et al., 2018; Choi and Lee, 2020).

TE transcripts that are unable to encode TE proteins may contribute to long non-coding RNAs (lncRNAs). It is reported that about 80% of the lncRNAs in *Arabidopsis* are polyadenylated (Di et al., 2014). While the majority of the lncRNAs are transcribed by RNA Pol II and acquire polyadenylated tails (Chekanova, 2015), some lncRNAs produced by plant-specific RNA polymerases, Pol IV and Pol V, are non-polyadenylated (Li et al., 2015; Wierzbicki et al., 2008). In spite of their important role in RdDM (see chapter 1), Pol IV- and Pol V-transcribed non-polyadenylated lncRNAs have been difficult to identify, possibly due to their extremely low abundance and poor stability (Chekanova, 2015), and are mostly produced from silenced TE loci rather than the active ones (Cuerda-Gil and Slotkin, 2016). A mixture of polyadenylated and non-polyadenylated RNAs would exponentially complicate the analysis strategy. This research, therefore, focuses on approaches based on Pol II-transcribed transcriptome data, as TE loci competent of autonomous transposition would have required Pol II-dependent transcription. In addition to the full-length and autonomous TE transcripts, polyadenylated transcripts associated with fragmented and non-autonomous TE loci are also of great importance in this study since the degenerated TE transcripts produced by Pol II may shed light on the relationship between transcriptionally active TE loci and neighbouring genes.

### 2.2.3 The pros and cons of existing tools for TE transcription analysis

Challenges arise when it comes to analysing TEs with second-generation (short-read) sequencing data. The characteristic repetitive and high copy number nature of TE loci are the prime cause of ambiguous alignment and assembly of such sequence data. Conventional strategies for mapping and quantification of transcriptome sequence data will discard multi-mapping reads or will uniformly divide a multi-mapping read to features having equally good alignment. For example, the frequently used software HTSeq-count (Anders et al., 2015) that was initially designed to quantify gene expression using only unique-mapping reads now has the option to count a multi-mapping read by equally scoring that read against all features it was assigned. Cufflinks (Roberts et al., 2011; Trapnell et al., 2010) uses the equal-weighting strategy and divides a multi-mapping read evenly to all of its mapped positions by default. It can also perform a re-estimation of read abundance by using the

abundance of unique-mapping reads as an initial probability distribution before re-assigning multi-mapping reads to the matched loci accordingly. Using this strategy, features having unique-mapping reads would be favoured over those obtain multi-mapping reads only. Nonetheless, these tools are not designed for the analysis of TE short read sequence data. Hence, most approaches utilise a 'masked' version of the reference genome data to avoid alignment aberrations.

To address these issues and allow sensible determination of the transcriptional activity of TEs in genomes, TEtranscripts (Jin et al., 2015) was developed. Specifically developed for TE analysis, this package applies the equal-weighting principle and estimates the "relative read abundance" (as the proportion to the read abundances of total TE loci) for each TE locus (please see Appendix A.1 for details). An expectation-maximization (EM) algorithm is then utilized to iteratively optimize the re-estimation of "relative read abundance" for each TE locus until the "relative read abundance" comes to convergence. This algorithm, however, might cause bias on closely related TE loci (e.g. TE loci of the same family), where TE loci that are shorter in length or initially assigned with more equal-weighting reads tend to be re-estimated with more "relative read abundance" (please see Appendix A.1 for details). This bias is likely to be minimized internally by this tool's last build-in computational step, which sums up the "relative read abundance" of individual TE loci at the family level. Nevertheless, the individual origins (TE loci) that generate the TE transcripts remain unknown.

With the incorporation of an equal weighting strategy and the EM algorithm, TEtranscripts provides a solution to mathematically better resolving the ambiguous alignments by considering the number of matched sites in the genome and the "effective length" of mapped features. The reality in the TE biology world, however, can be much more complicated. It may need an even more complicated computational model to cover and evaluate all possible alignment scenarios. Alternatively, the combination of multiple approaches, each depicting part of the TE transcription nature, might widen the ability to identify transcriptionally active TE loci.

### 2.2.4 A plan to establish a workflow combining existing tools

As described above, each set of tools has a narrow ability to describe TE transcriptional activity. Individually all fall short of being able to unambiguously assign transcription to specific loci – a key to the identification of potentially autonomous elements that represent active elements within genomes. We initially posed the question of whether using a combination of existing approaches might be more effective at identifying loci that house active, autonomous elements.

Although a TE family can proliferate through the genome generating hundreds or thousands of insertions, each individual loci accumulates single nucleotide variants (SNVs) and small insertions and deletions (INDELs) over time. These mutations, depending on location and volume within an

individual element, can affect a TE's competency for autonomous mobilization to various degrees. While some mutations erode a TE's ability for self-proliferation, others have a trivial influence on competent transposition thus can be inherited to new TE insertion sites until a newly emerged mutation jeopardize the mobilization competency (Huang et al., 2012). These mutations exist as pedigree traces of TE dynamics and underlie the formation of sub-lineages within a TE family. While some sub-lineages contain transcriptionally active TE individuals, others might be entirely silenced. Therefore, these mutations might facilitate the positive identification of a transcriptionally active TE locus or TE sub-lineage within a family containing thousands of loci, most of which might not be transcribed. So rather than developing a new computational model to resolve the alignment ambiguity, to distribute multi-mapping reads to distinct loci, or to quantify the expression level of individual TEs, this chapter aims to establish a pipeline that utilises existing tools, each identifying a subset of transcribed elements, to collect all potentially expressed TE loci.

The approach that we designed firstly uses the default function of HTSeq-count (Anders et al., 2015) to collect transcriptionally active TEs having unique reads. Secondly, to broadly capture TEs aligned with unique- or multi-mapping reads, bedtools coverage and bedtools intersect (Quinlan and Hall, 2010) were used collectively. In this part, each equally good alignment of a multi-mapping read would be counted as one hit without further weighting process. Thirdly, with paired-end sequencing, read pairs mapping across TE boundaries can be leveraged for capturing TE loci obtaining multi-mapping reads only within the TE feature, yet the read mates align to a unique location in the fragment size range of the sequenced library. The software TEFingerprint (Plant and Food Research, 2019) was developed for this purpose for analysing DNA sequencing data, while this research intends to explore its ability on transcriptome data. More details for this pipeline and approach are described in Materials and Methods, section 2.3 below.

This chapter, therefore, focuses on the establishment of a new analysis pipeline to identify a reduced pool of potentially transcriptionally active TE loci. The following chapter (chapter 3) describes the use of this pipeline to determine the change in the transcriptional landscape of transposons in grapevine embryogenic callus exposed to a range of physical and pathogen stressors.

## 2.3 Methods

### 2.3.1 Stress treatment

Embryogenic callus cultures were established from *Vitis vinifera* cv *Pinor noir* clone UCD5 and maintained according to Lizamore (2013). Stress treatment was conducted essentially based on the methods established by Lizamore (2013). Every half gram of solid embryogenic callus cultures was collected into a 15-mL Corning conical polypropylene centrifuge tube (Sigma) containing 12 mL of hormone-free $C1^P$ liquid medium (HF- $C1^P$; see Appendix B for recipe) that have been supplied with either live *Hanseniaspora uvarum* cultures resuspended to $OD_{600}$ = 0.8 or extracts of *Botrytis cinerea* as described in Lizamore (2013). These embryogenic callus cultures exposed to stressors were subjected to vigorous shaking for 2 seconds manually, following with incubation on a rotary shaker (100 RPM) horizontally for 8 minutes at room temperature. These calli were then transferred and spread evenly on HF- $C1^P$-soaked filter papers on fresh $C1^P$ plates. Continuously with the presence of the biotic stressors, these plates were incubated at 25˚C in the dark across a time series as described in Figure 2.1. Preparation of *H. uvarum* yeast culture and *B. cinerea* cell extracts was carried out as described in Lizamore (2013). Mock experiments were identical to that carried out for treatments with live yeast cultures or fungal extracts with same volume of stressor-free HF- $C1^P$ liquid medium. Each treatment was harvested at 1, 3, 6, and 12 hours as shown in Figure 2.1, and gently washed with 50mL of HF- $C1^P$ three times before being snap-frozen with liquid nitrogen. A common untreated 0 hour time point (denoted as T=0) for mock and the two biotic treatments was taken prior to any treatment. All treatments and their associated time points consisted of three biological replicates.

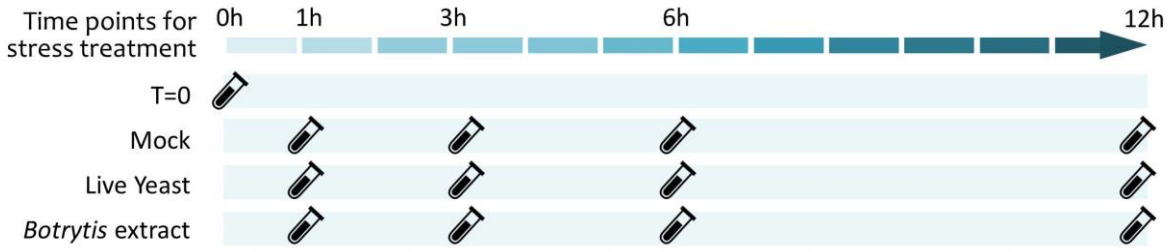



**Figure 2.1 Experimental settings of stress treatment**

### 2.3.2 RNA-seq library preparation and sequencing

Total RNA of harvested embryogenic callus was isolated by Tirthanker Gosh according to the manufacturer's instructions of Spectrum Plant Total RNA Kit (Sigma). Each purified RNA sample was

treated with DNase I following the protocol of TURBO DNA-free kit (Ambion) to remove contaminating genomic DNA before being sent to the then New Zealand Genomics Ltd (now Otago Genomics) for library preparation and pair-end Illumina sequencing using a HiSeq 2500 sequencer.

### 2.3.3 Annotation of *V. vinifera* TEs

A new and more complete TE annotation file was established by Lizamore (2013). In general, all canonical transposable element sequences of *V. vinifera* were downloaded from the Repbase Update database. In order to reconstruct the full canonical element of LTR-retrotransposons (LTR-TEs), LTRs and the internal sequences were reassembled for each family in the format of LTR-Internal-LTR. The canonical TE sequences were used to extract local copies of the 12X PN40024 grapevine genome (Ensembl Plants database) using RepeatMasker with the default setting. Simple sequence repeats (SSRs) were masked with N's using RepeatMasker. In total, 232 canonical TE sequences, in which each sequence represents a TE family of *V. vinifera*, were used to extract 223,411 TE-like sequences of the 12X PN40024 genome.

### 2.3.4 Bioinformatics analysis

#### RNA-seq data preprocessing

Adapter sequences and low-quality bases were trimmed by fastq-mcf (Aronesty, 2013) before quality check using fastqc (Andrews, 2010). Trimmed quality reads were aligned to *V. vinifera* tRNA and rRNA sequences (Ensembl Plants database), followed by the collection of unmapped reads using samtools (Li et al., 2009). The detail of all bioinformatics commends and scripts used in pre-processing are provided in Appendix D.1.

#### Sequence mapping and cut-off threshold for expressed genes and TEs

Sequencing reads unmapped to grapevine's tRNA and rRNA sequences were aligned to 12X PN40024 grapevine reference genome using HISAT2 (Kim et al., 2015a) with the following parameters: `–rna–stradness RF –dtk –k 100`. Reads mapping to genes were then quantified by htseq-count with parameters `-f bam -t exon -i transcript_id -s reverse -m intersection-nonempty`. Genes' GTF file (version v2.1) was provided by downloaded from the Grape Genome Database hosted at CRIBI (http://genomes.cribi.unipd.it/grape/) on 19th July 2016. Fragments per kilobase of gene model per million mapped reads (FPKM) values were calculated per gene with R scripts.

A common cut-off threshold for expressed genes was set at FPKM > 1. Due to the repetitive characteristics and high similarity shared among TEs, however, most TE-related reads were multi-

mapping reads that cannot be used directly for FPKM estimation. Even if a TE obtained unique-mapping reads, these reads would only map to a very small proportion of that TE; thus, TE-related unique-mapping reads are not suitable for FPKM calculation either. Therefore, a general cut-off threshold applied on raw read count of expressed TEs for the 39 sequencing libraries was inferred from normalized FPKM (zFPKM) of genes according to Hart et al. (Hart et al., 2013). Genes with FPKM=0 were excluded before transforming FPKM into zFPKM using the Bioconductor suite zFPKM (Ammar and Thompson, 2019). Given the recommended threshold of zFPKM >= -3 (Hart et al., 2013), genes with zFPKM values that fell between -2.99 and -3 were collected to estimate the FPKM threshold across the 39 libraries. In this case, the pooled raw count per 3kb transcript from genes having zFPKM between -2.99 and -3 was around 10, concordant with the observations in Hart et al. (2013) and with the minimum read count recommended for differential analysis (Soneson and Delorenzi, 2013). As a result, the common raw count ten inferred from the zFPKM cut-off boundary was used as the general cut-off threshold of the pipeline described in the following section for transcriptionally active TEs.

## Analysis pipeline for collecting TE expression candidates

### *Sub-pipeline 1*

This sub-pipeline only collects TEs obtaining unique-mapping reads that each of these sequencing reads can be traced back to a unique origin in the genome. As illustrated in Figure 2.2A, sequencing reads unmapped to grapevine's tRNA and rRNA sequences were aligned to 12X PN40024 grapevine reference genome using HISAT2 (Kim et al., 2015a) with the parameters: `–rna-stradness RF –dtk –k 100`. Reads mapping to TEs were then quantified by htseq-count (Anders et al., 2015), in which only uniquely mapped reads were counted. TEs having read count more than 10, which approximates to 5 pairs of read, were collected.

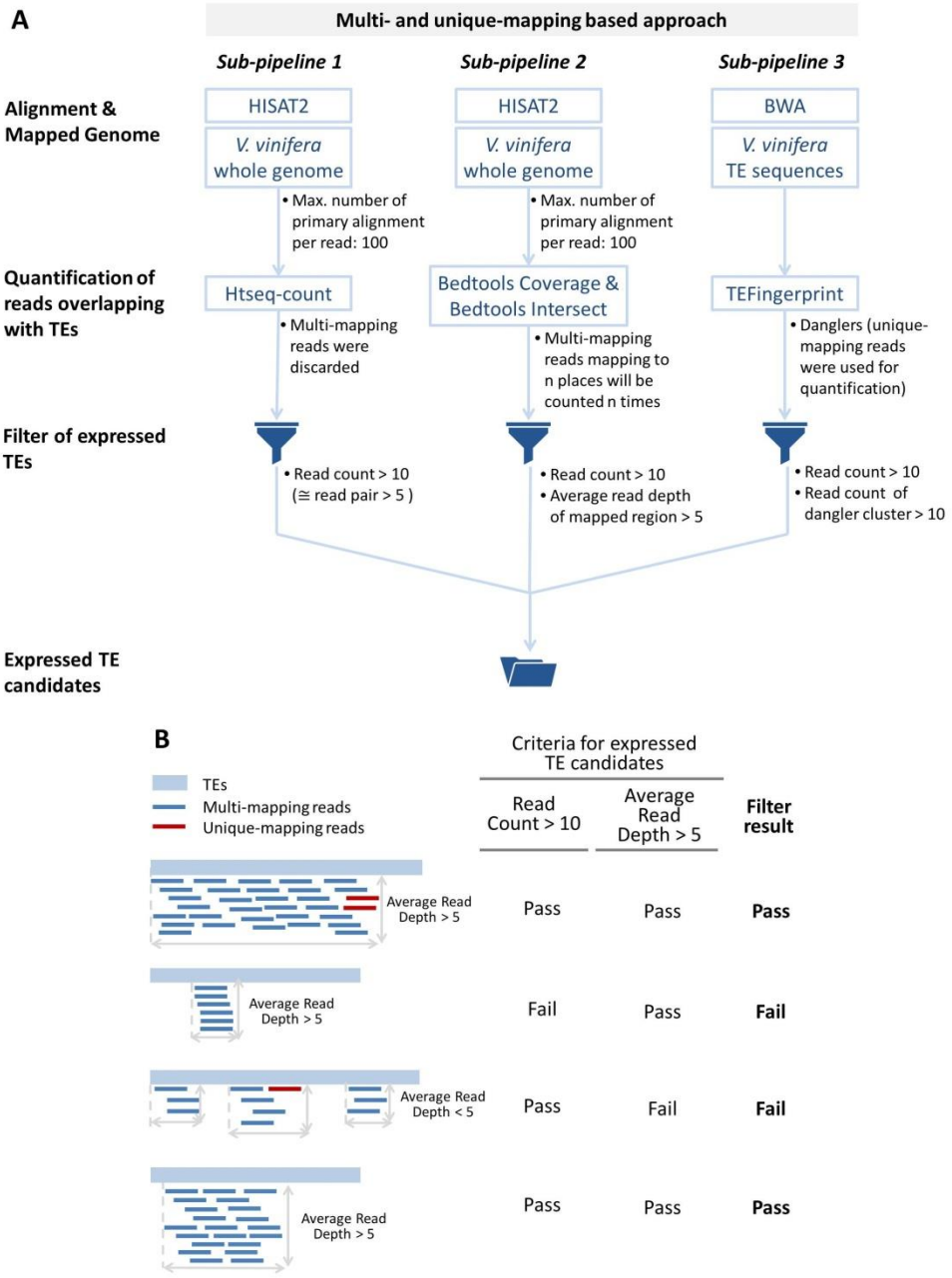



**Figure 2.2 The pipeline of identification of expressed TE candidates**

**(A)** The first and second sub-pipelines apply HISAT2 for the alignment of sequencing reads against the reference genome and then use htseq-count and BEDtools tool set, respectively, to quantify reads overlapping with TEs. While Htseq-count only adopts unique-mapping reads, BEDtools incorporates both unique- and multi-mapping reads. The third sub-pipeline uses BWA to align reads against TE sequences, after which the mates of TE-mapped reads would be fed into TEFingerprint for mapping against the reference genome to capture danglers. The three sets of TEs passing through the filtering steps are joined together as a pool of expression candidates. **(B)** Examples of filter step for the BEDtools-based sub-pipeline. To pass this step, a TE needs to show > 10 read count and >5 average read depth normalized by TE's mapped region.

***Sub-pipeline 2***

The concept of this sub-pipeline was to collect individual TEs that were aligned with any kind of reads, irrespective of the number of highest quality mapping loci for a given read. Following read alignment using HISAT2 as described in sub-pipeline 1, the BEDtools suite (Quinlan and Hall, 2010) was used in read quantification (Figure 2.2A, sub-pipeline 2). The command `bedtools coverage` generated raw count for TEs while multi-mapping reads matching to n-places (e.g. 10) was recorded n-times (i.e. 10). It also counted the number of bases of a TE locus covered by reads (covered bases of TE). Furthermore, this sub-pipeline incorporated `bedtools intersect` to calculate the number of bases of a read overlapping with an individual TE (mapped bases of read), which were summed for each TE locus to estimate the average read depth of an individual TE's mapped region (i.e. only the region covered by reads, not the entire annotated feature). This was calculated as follows:

For $n$ reads mapping to a TE locus, and $i$ as an integer from 0 to $n$, $f(i) =$ mapped bases of read $i$.

Average read depth of an individual TE's mapped region = $\frac{\sum_{i=0}^{n} f(i)}{\text{Bases of a TE locus covered by reads}}$ .

In order to exclude TE loci that were covered by reads in a sparse and scattered way, a cut-off threshold of the average read depth 5 was adopted in addition to the 10 read count threshold. Examples of the filtering step of this sub-pipeline were illustrated in Figure 2.2B.

***Sub-pipeline 3***

The third part (Figure 2.2A, sub-pipeline 3) specifically collects TEs had transcription across the boundaries of the element. The in-house software TEFingerprint was originally designed for identifying unannotated insertions in genomes using paired-end short fragment DNA sequence data. Here it was applied to capture TE loci internally mapped by multi-mapping reads only, yet the read mates, as known as danglers, were uniquely aligned to a location near the insertion site (Figure 2.3). Although TEFingerprint also has the option to use reads sit across the junction of a TE locus, this function was disabled in sub-pipeline 3 as the utilization of htseq-count in sub-pipeline1 has covered this scenario. Sub-pipeline 3 applied the standard TEFingerprint pipeline where reads were mapped against the collection of 223,411 annotated *V. vinifera* TE sequences using BWA (Li and Durbin, 2009). Subsequently, the mates of TE-mapped reads were aligned to the reference genome (12X PN40024) before calculating the read count of dangler clusters. Only clusters containing more than 10 dangler reads were kept to test for the intersection of dangler clusters and annotated TEs using `bedtools intersect`. The candidates need to show more than 10 dangler reads and more than 10 reads mapping internally (counted by `bedtools coverage`).

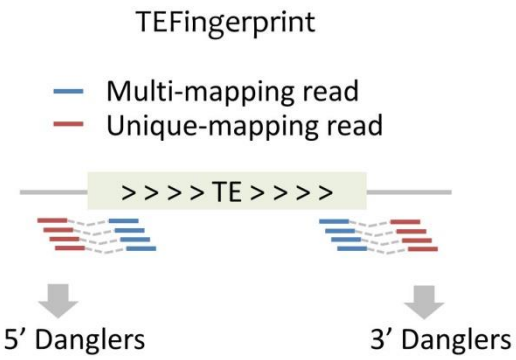



**Figure 2.3 Illustration of the mapping strategy of TEFingerprint used in the pipeline**

The blue fragments denote the reads mapping to TE sequences, whereas the red ones are the corresponding read mates (danglers) of TE-mapped reads, which are indicative of the transcriptional activity of a TE with transcription across boundaries.

***Summarizing TE expression candidates***

After excluding TEs did not show enough signal of transcription, potentially expressed TEs from the three sub-pipelines were collected together as a pool of expression candidates.

## TEtranscripts analysis

As described in 3.3.4.1 and 3.3.4.2, the raw sequencing reads were trimmed and reads mapping to tRNA and rRNA were removed before mapping the remaining reads to the reference genome. The resulting BAM files sorted by position were passed to TEtranscripts (Jin et al., 2015) with the following arguments: `--sortByPos --mode multi --stranded reverse`. The cut-off threshold for expressed TE family was set at 10.

All computational scripts used in this chapter can be found in Appendix D.1.

## 2.4 Results

### 2.4.1 Alignment statistics

There are 223,411 TE-related DNA elements (loci) in the grapevine genome corresponding with 232 canonical TE sequences, each representing a TE family, deposited in Repbase (Table 2.1). These elements occupy about 33% of the 500 Mb-genome. To comprehensively survey transcriptional activity of TEs in Pinot noir embryogenic callus cultures, short sequencing reads of stranded polyadenylated transcriptome data were analyzed with the pipeline incorporating unique- and multi-mapping reads (Figure 2.2). Basic alignment statistics were obtained at multiple stages of the preprocessing and alignment procedures with the aligner HISAT2. The sequencing produced 32 to 46 million reads per library (Table 2.2). Twenty-seven of 39 libraries showed alignment rate above 80%, and seven libraries had alignment rates between 70% and 80%. Four libraries, mostly from the 1 and 3 hours of yeast treatment, only had 60% to 70% of total sequencing reads mapped to grapevine's reference genome, and one library of 3-hour yeast treatment obtained just 44% of sequencing reads mapped. The abundance of total sequenced reads and reads passed quality check `was similar across all libraries. The various proportions of reads mapping to tRNA or rRNA might partially contribute to the variation of alignment rates, as the comparison between the number of mapped reads and number of reads after removing tRNA/rRNA-mapped reads adjusted the alignment rates to 70-90% for most of the samples (Table 2.2). Although the alignment rate of the third replicate of the 3-hour yeast treatment was still relatively low compared to other libraries, it was lifted from 44% to 57% after the adjustment. It was speculated that the considerable proportion of unmapped reads in this library might be related to the inevitable contamination of yeast cells in the collected grapevine embryogenic callus cultures. Re-alignment of the unmapped reads against either *Saccharomyces Cerevisiae* or *H. uvarum* reference genomes (both acquired from NCBI) revealed that 1.7%, 2.7%, 29%, and 35% of the unmapped reads, respectively, in 1, 3, 6, and 12 hours of yeast treatment were specifically mapped to *H. uvarum* (Appendix C.1). It seemed that the contamination of yeast mRNA in the 3-hour yeast treatment could not fully explain the relative low alignment rates to the grapevine genome. Because our analysis pipeline focuses on the identification of transcriptionally active TEs instead of quantifying expression level, and our experimental design includes three technical replicates for each time-point, we reckoned that this level of variation in alignment rates was acceptable. Besides, the utilization of FPKM (Fragments Per Kilobase of transcript per Million mapped reads) can normalize read counts by incorporating the abundances of mapped reads, and tools for differential expression analysis, like DESeq2 (Love et al., 2014), can normalize raw read counts by stabilizing the variances across libraries before performing statistical test, therefore we decided to keep this library for further analysis in chapter 3 and chapter 4.

**Table 2.1 Annotation of *V. vinifera* TE loci based on the canonical TE sequences extracted from the *V. vinifera* division in Repbase.**

| Class | Subclass | Superfamily | # TE families | # TE loci | Elements percentage (%) | Sequence occupied (bp) | Sequence percentage of all TE loci (%) | Sequence percentage of genome (%) |
|---|---|---|---|---|---|---|---|---|
| **Type I** | **Retrovirus-like** | ***Caulimorvirus*** | 3 | 3,056 | 1.37 | 2,597,652 | 1.62 | 0.53 |
| | **LTR** | ***Copia*** | 101 | 44,598 | 19.96 | 39,832,291 | 24.83 | 8.19 |
| | | ***Gypsy*** | 36 | 64,827 | 29.02 | 67,153,758 | 41.86 | 13.81 |
| | **non-LTR** | ***LINE*** | 10 | 23,447 | 10.50 | 19,410,035 | 12.10 | 3.99 |
| **Type II** | **TIR** | ***CACTA*** | 15 | 12,632 | 5.65 | 5,555,042 | 3.46 | 1.14 |
| | | ***Harbinger*** | 11 | 32,053 | 14.35 | 7,716,768 | 4.81 | 1.59 |
| | | ***hAT*** | 15 | 15,374 | 6.88 | 6,132,240 | 3.82 | 1.26 |
| | | ***MULE*** | 40 | 27,336 | 12.24 | 11,965,668 | 7.46 | 2.46 |
| | **non-TIR** | ***Helitron*** | 1 | 88 | 0.04 | 66,952 | 0.04 | 0.01 |
| **TOTALS** | | | **232** | **223,411** | **100.00** | **160,430,406** | **100.00** | **33.00** |

**Table 2.2 Mapping statistics for RNA-seq.**

| Sequenced libraries | | | Sequenced reads | | Adaptor removal & quality check | | Filter tRNA/rRNA | | Mapped reads | | Mapped / tRNA rRNA filtered |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Treatments | Timepoints | Replicates | | | | | | | | | |
| Control | 00 h | a | 44,843,044 | 100% | 44,843,012 | 100% | 44,692,686 | 99.66% | 40,239,959 | 89.74% | 87.80% |
| | | b | 44,755,712 | 100% | 44,755,674 | 100% | 44,596,226 | 99.64% | 40,082,495 | 89.56% | 89.24% |
| | | c | 42,000,138 | 100% | 42,000,108 | 100% | 41,856,356 | 99.66% | 37,141,431 | 88.43% | 88.46% |
| Mock | 01 h | a | 42,191,300 | 100% | 42,191,278 | 100% | 41,962,262 | 99.46% | 36,842,403 | 87.32% | 87.80% |
| | | b | 42,223,790 | 100% | 42,223,772 | 100% | 42,001,096 | 99.47% | 37,483,039 | 88.77% | 89.24% |
| | | c | 43,672,550 | 100% | 43,672,524 | 100% | 43,403,184 | 99.38% | 38,393,130 | 87.91% | 88.46% |
| | 03h | a | 43,786,332 | 100% | 43,786,314 | 100% | 43,493,816 | 99.33% | 38,843,464 | 88.71% | 89.31% |
| | | b | 44,152,592 | 100% | 44,152,550 | 100% | 43,928,442 | 99.49% | 38,941,198 | 88.20% | 88.65% |
| | | c | 45,385,724 | 100% | 45,385,706 | 100% | 45,197,900 | 99.59% | 40,702,357 | 89.68% | 90.05% |
| | 06 h | a | 41,021,758 | 100% | 41,021,730 | 100% | 40,768,358 | 99.38% | 36,341,402 | 88.59% | 89.14% |
| | | b | 42,305,244 | 100% | 42,305,222 | 100% | 42,056,918 | 99.41% | 37,187,926 | 87.90% | 88.42% |
| | | c | 44,026,700 | 100% | 44,026,676 | 100% | 43,769,928 | 99.42% | 38,917,371 | 88.39% | 88.91% |
| | 12 h | a | 42,102,864 | 100% | 42,102,834 | 100% | 41,988,512 | 99.73% | 37,184,585 | 88.32% | 88.56% |
| | | b | 45,386,716 | 100% | 45,386,692 | 100% | 45,253,416 | 99.71% | 39,938,420 | 88.00% | 88.26% |
| | | c | 43,324,658 | 100% | 43,324,636 | 100% | 43,188,460 | 99.69% | 38,240,351 | 88.26% | 88.54% |
| Yeast | 01 h | a | 43,575,622 | 100% | 43,575,594 | 100% | 38,285,868 | 87.86% | 28,234,233 | 64.79% | 73.75% |
| | | b | 44,625,058 | 100% | 44,625,030 | 100% | 40,134,482 | 89.94% | 31,176,032 | 69.86% | 77.68% |
| | | c | 44,487,058 | 100% | 44,487,028 | 100% | 41,552,174 | 93.40% | 33,952,493 | 76.32% | 81.71% |
| | 03h | a | 42,808,896 | 100% | 42,808,872 | 100% | 39,022,006 | 91.15% | 30,551,032 | 71.37% | 78.29% |
| | | b | 41,787,532 | 100% | 41,787,514 | 100% | 36,478,946 | 87.30% | 26,317,172 | 62.98% | 72.14% |
| | | c | 42,671,442 | 100% | 42,671,420 | 100% | 32,940,188 | 77.19% | 18,842,818 | 44.16% | 57.20% |
| | 06 h | a | 43,203,508 | 100% | 43,203,490 | 100% | 43,036,378 | 99.61% | 36,694,381 | 84.93% | 85.26% |
| | | b | 44,381,190 | 100% | 44,381,164 | 100% | 43,965,004 | 99.06% | 36,890,500 | 83.12% | 83.91% |
| | | c | 41,684,294 | 100% | 41,684,266 | 100% | 40,522,008 | 97.21% | 32,962,657 | 79.08% | 81.35% |
| | 12 h | a | 41,676,868 | 100% | 41,676,826 | 100% | 40,466,602 | 97.10% | 33,085,049 | 79.38% | 81.76% |
| | | b | 41,697,894 | 100% | 41,697,860 | 100% | 40,655,260 | 97.50% | 36,804,651 | 88.27% | 90.53% |
| | | c | 42,781,030 | 100% | 42,781,002 | 100% | 41,924,084 | 98.00% | 34,470,211 | 80.57% | 82.22% |
| Botrytis | 01 h | a | 41,249,900 | 100% | 41,249,884 | 100% | 37,472,014 | 90.84% | 29,488,924 | 71.49% | 78.70% |
| | | b | 41,059,148 | 100% | 41,059,120 | 100% | 40,851,020 | 99.49% | 36,081,733 | 87.88% | 88.33% |
| | | c | 43,246,768 | 100% | 43,246,732 | 100% | 38,874,426 | 89.89% | 29,804,813 | 68.92% | 76.67% |
| | 03h | a | 44,129,956 | 100% | 44,129,910 | 100% | 41,940,624 | 95.04% | 35,009,493 | 79.33% | 83.47% |
| | | b | 44,035,740 | 100% | 44,035,698 | 100% | 41,130,150 | 93.40% | 33,756,289 | 76.66% | 82.07% |
| | | c | 44,492,730 | 100% | 44,492,696 | 100% | 42,883,184 | 96.38% | 36,783,352 | 82.67% | 85.78% |
| | 06 h | a | 44,283,046 | 100% | 44,283,010 | 100% | 43,425,150 | 98.06% | 37,252,117 | 84.12% | 85.78% |
| | | b | 39,970,310 | 100% | 39,970,286 | 100% | 39,518,750 | 98.87% | 34,770,797 | 86.99% | 87.99% |
| | | c | 40,814,706 | 100% | 40,814,678 | 100% | 40,278,752 | 98.69% | 34,907,361 | 85.53% | 86.66% |
| | 12 h | a | 41,096,644 | 100% | 41,096,610 | 100% | 40,165,918 | 97.74% | 34,925,753 | 84.98% | 86.95% |
| | | b | 32,464,438 | 100% | 32,464,420 | 100% | 31,585,116 | 97.29% | 27,274,821 | 84.01% | 86.35% |
| | | C | 43,649,536 | 100% | 43,649,518 | 100% | 42,489,256 | 97.34% | 36,674,118 | 84.02% | 86.31% |

Given the ubiquitous distribution of TEs, TE-related reads contributed to less than 2.5% of the total mapped reads across libraries (Figure 2.4A). The paired-end sequencing allows the distinction between sense and anti-sense alignment, which is crucial for TE analysis. Although most of the TE-related reads aligned with TEs in sense orientation, about 40% of TE-mapped reads were derived from transcripts antisense to the orientation of mapped TEs (Figure 2.4B). Note that a very small

proportion of TE-mapped reads were found to map in both sense and antisense orientations. These reads can be categorized into one of the three scenarios (Figure 2.5A). The first represents read mapped to at least two closely co-localized TEs on different strands (Figure 2.5A, category 1). The second category denotes read not only mapped to at least two adjacent TEs in different orientations but also mapped to another TE in an antisense orientation (Figure 2.5A, category 2). The last scenario depicts read not only mapped to one TE in sense orientation but also mapped to another TE as antisense transcript (Figure 2.5A, category 3). Over 97% of reads with dual alignment behaviours were grouped into category 1, whereas category 2 is the second largest group (Figure 2.5B). Category 3 only contributed to less than 1% of this read subset.

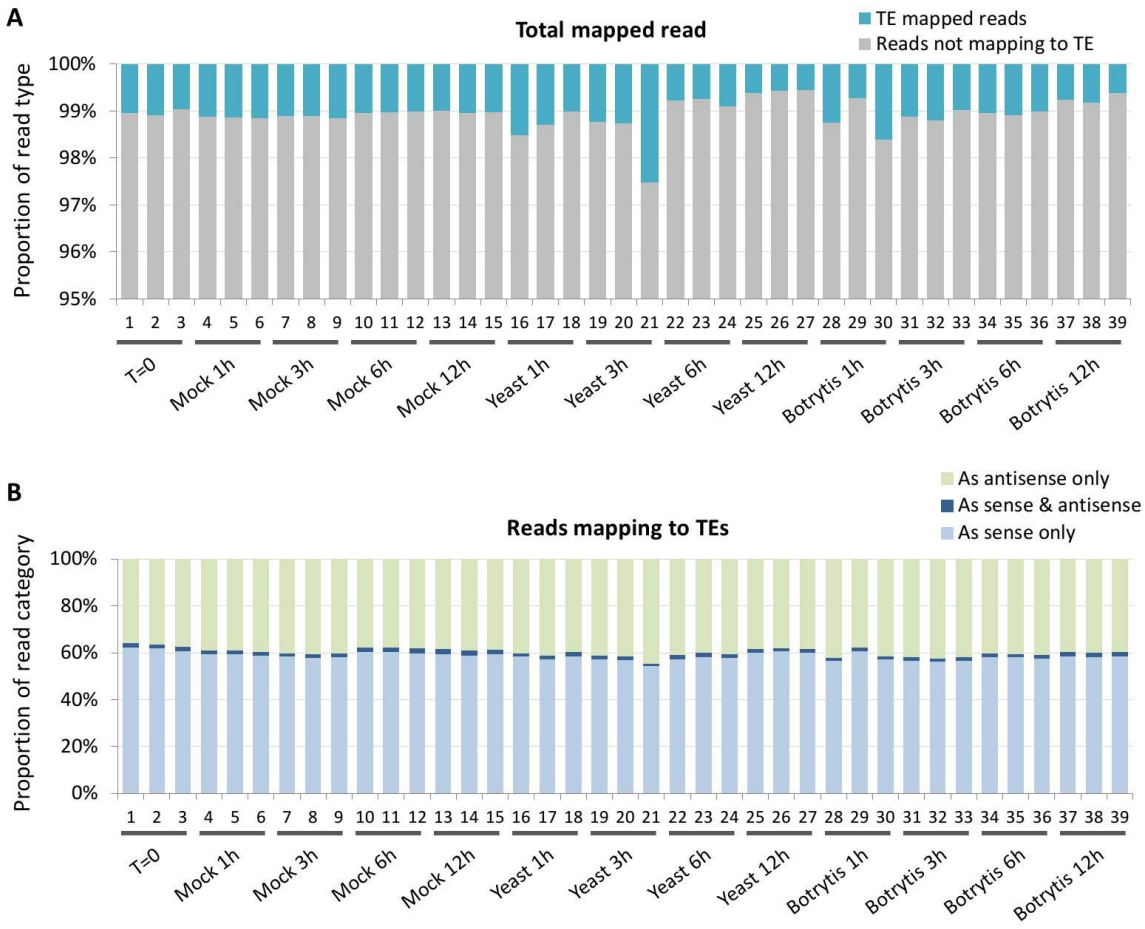



**Figure 2.4 Mapping statistics of total mapped reads and TE-related reads**

**(A)** Proportion of TE-related reads of each sequencing library were highlighted in blue colour. Due to the small proportion of TE-mapped reads, the Y axis only shows the range from 95% to 100% **(B)** TE-related reads can be grouped into 3 categories according to their alignment behavior: sense (light blue), antisense (light green), or dual (dark blue).

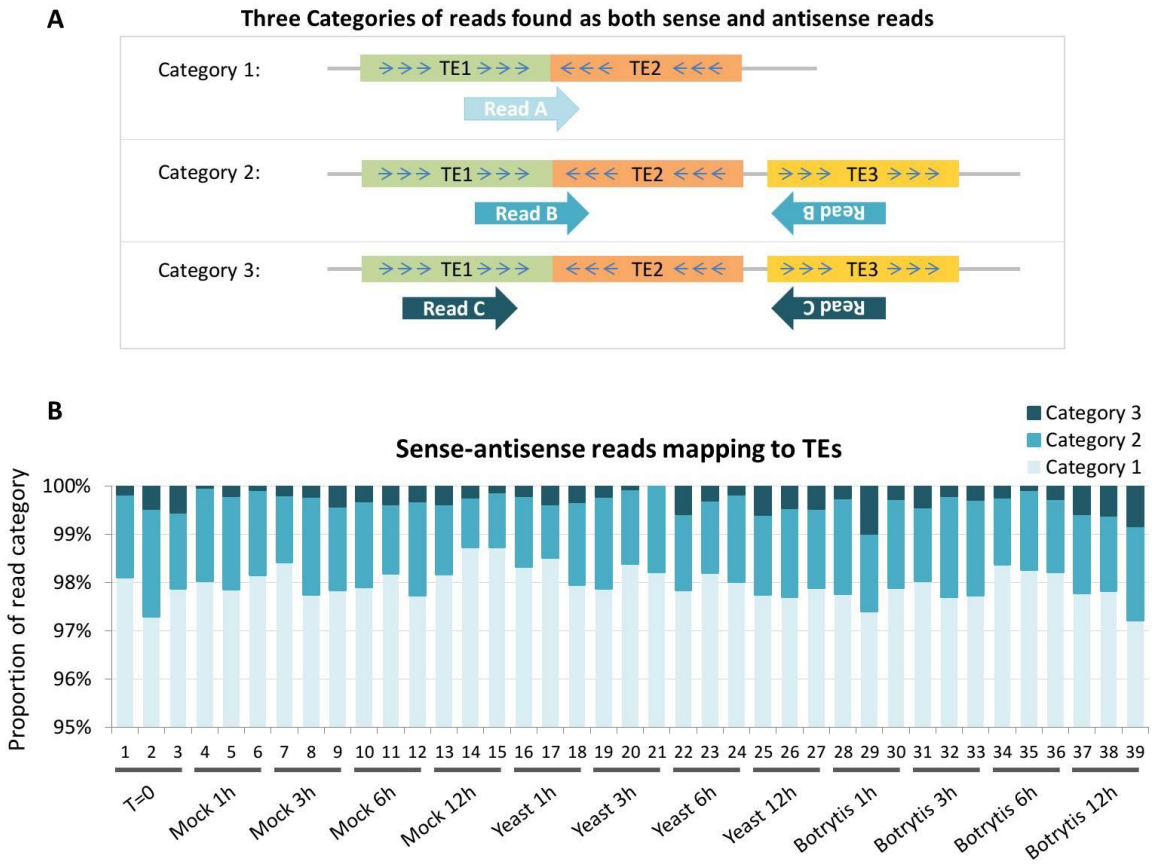



**Figure 2.5 Alignment scenarios of TE-related reads showing dual mapping manner**

**(A)** TE-related reads with dual mapping behaviour can be assigned into one of the three scenarios. While the first one represents read mapping to two adjacent but different oriented TEs, sequencing read in the second case also maps another TE in the antisense direction. The third scenario demonstrates read mapping to at least two distant TEs in distinct orientations. **(B)** The proportion of read in different scenarios for each library. Only the proportion range between 95% and 100% were shown to reveal the details.

### 2.4.2 Application of the TE expression candidate analysis pipeline

The pipeline for the collection of potentially expressed TE individuals was applied on the RNA-seq data generated from *P. noir* UCD5 embryogenic callus treated with live yeast or *Botrytis* extract over multiple time points. For embryogenic callus cultures at T=0, this pipeline excluded 87% of annotated TE loci that had no mapping reads at all (Figure 2.6A). It also excluded about 11% TEs that were sparsely and randomly covered by sequencing reads thus fell under the threshold. In other words, only 1.6% of TEs are potentially expressed. For the RNA-seq data derived from mock, yeast and Botrytis treatments, the pipeline had also effectively excluded 75% to 80% of TE individuals as having no evidence of expression (Figure 2.6B-D), whereas 17% to 22% of TEs were excluded due to unsufficient support of read count and average read depth normalized by TE mapped region. Hence over 97% of TE loci were not considered to be valid expression candidates. Thus the remaining 2.3% to 2.5% of annotated TE loci across all treatments, equivalent to 5,170 to 5,530 TE individuals, were

considered to represent the pool of expression candidates. The expression candidates found in each treatment accounted for 70% to 85% of the reads mapping to TEs in sense orientation in the corresponding libraries (Figure 2.7), although a small proportion (< 5%) of reads mapping to both expression candidates and non-candidates were observed in each library.

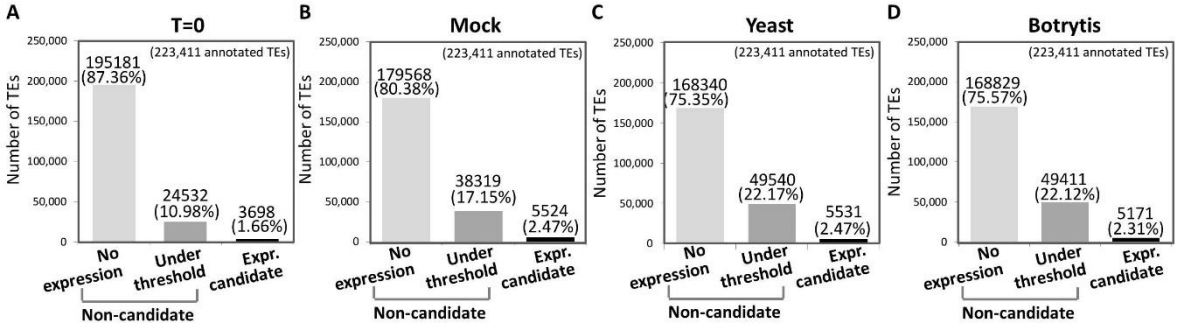



**Figure 2.6 Expression candidates identified by the pipeline across various treatments**

All annotated TEs were categorized by transcriptional activity indicated at the x-axis and illustrated according to treatments as shown in each graph. **(A)** T=0, **(B)** mock, **(C)** yeast treatment, **(D)** *Botrytis* treatment.

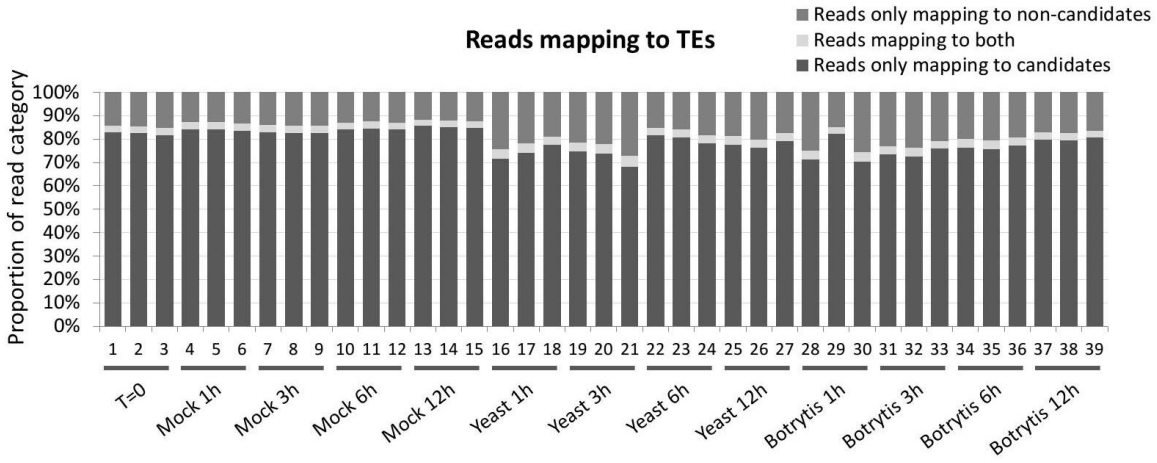



**Figure 2.7 Proportion of reads contributed from expression candidates and non-candidates**

Reads mapping to TEs in sense orientation were categorized into three categories, only mapping to expression candidates, only mapping to non-candidates, and mapping to both. Sequencing reads of each library were analyzed according to the pools of expression candidates and non-candidates captured in the corresponding stress treatments.

The exclusion of TE loci having zero mapped read is clearly sensible. Contrarily, there may not be an explicit boundary of read abundances or read depth to separate TE loci transcribed or sequenced at background noise level (i.e. under-threshold loci) from the expression candidates; thus, the suitability of the cut-off threshold needs to be carefully examined. Although the rationale for the cut-off threshold has been explained in section 2.3.4, the distribution of TE loci expression level was not clear. If the cut-off threshold sits at or close to a peak of expression level, any slight shift of the cut-off threshold might dramatically change the number of TE loci qualified as expression candidates, and therefore this scenario should be avoided. To compare the expression level of under-threshold TEs and expression candidates, and to visualize the position of cut-off threshold, individual TE's expression level, in terms of read abundances (including unique- and multi-mapping reads) and the average depth of mapped region acquired from BEDtools suites, was investigated. In this analysis, data of T=0 cultures was used as a representative of all libraries. Excluding TE loci that had zero mapped reads, the average read count from the three technical replicates for under threshold TE loci ranged from just above 0 to about 34,000 (x-axes of Figure 2.8A, B) and peaked at the vicinity between 0 and 2 reads. On the other hand, the average read depth normalized with the size of the mapped region (i.e. average depth of mapped region) ranged from above 0 to 1,451 and showed peaks between 0 and 2 average read depth (y-axes of Figure 2.8B, C). Under-threshold TE loci were coloured in grey, and the expression candidates captured by one or multiple sub-pipelines were shown with different colours in Figure 2.8B. The red dashed lines indicate the expression threshold for the BEDtools sub-pipeline placed at 10 reads and an average read depth of 5 (red dashed lines in Figure 2.8). The upper-right corner of Figure 2.8B was enriched with expression candidates that fulfilled the thresholds of all three sub-pipelines (red dots in Figure 2.8B), meaning TE loci with a higher level of read abundances and the average depth of mapped region would be more likely to be supported by all three methods. As the expression level moves downward, the expression candidates tend to be picked up by only one or two sub-pipelines (light green, yellow, dark green, and pink dots in Figure 2.8B). Note that some of the TE loci excluded in the BEDtools sub-pipeline due to low read depth were picked up by the Htseq sub-pipeline for their having over 10 unique-mapping reads or 5 unique-mapping read pairs (light blue dots in Figure 2.8B). Since the presence of unique-mapping reads is frequently considered as evidence of transcription, these TE loci were kept in the pool of expression candidates. Last but not least, the lower-left corner of Figure 2.8B was populated with under-threshold TE loci (grey dots) that were expressed as ambient background noise, which peaked between 0 and 2 reads (Figure 2.8A) or between 0 and 2 average read depth (Figure 2.8C). Altogether, this analysis revealed that the majority of the under-threshold TE loci were sparsely aligned by reads at an extremely low level away from the thresholds, which fulfil the expectation described above. This analysis also showed that the three sub-pipelines could together identify TE

loci potentially with high expression level (i.e. red dots in Figure 2.8B) and recognize loci with low to medium expression levels (i.e. dots with other colours except for grey in Figure 2.8B). A single application of any one of the sub-pipeline would have lost some of the expression candidates, whereas the combination of the three strategies widens the ability to include TE loci with unique- and multi-mapping reads and read-pairs mapping across TE's junction.

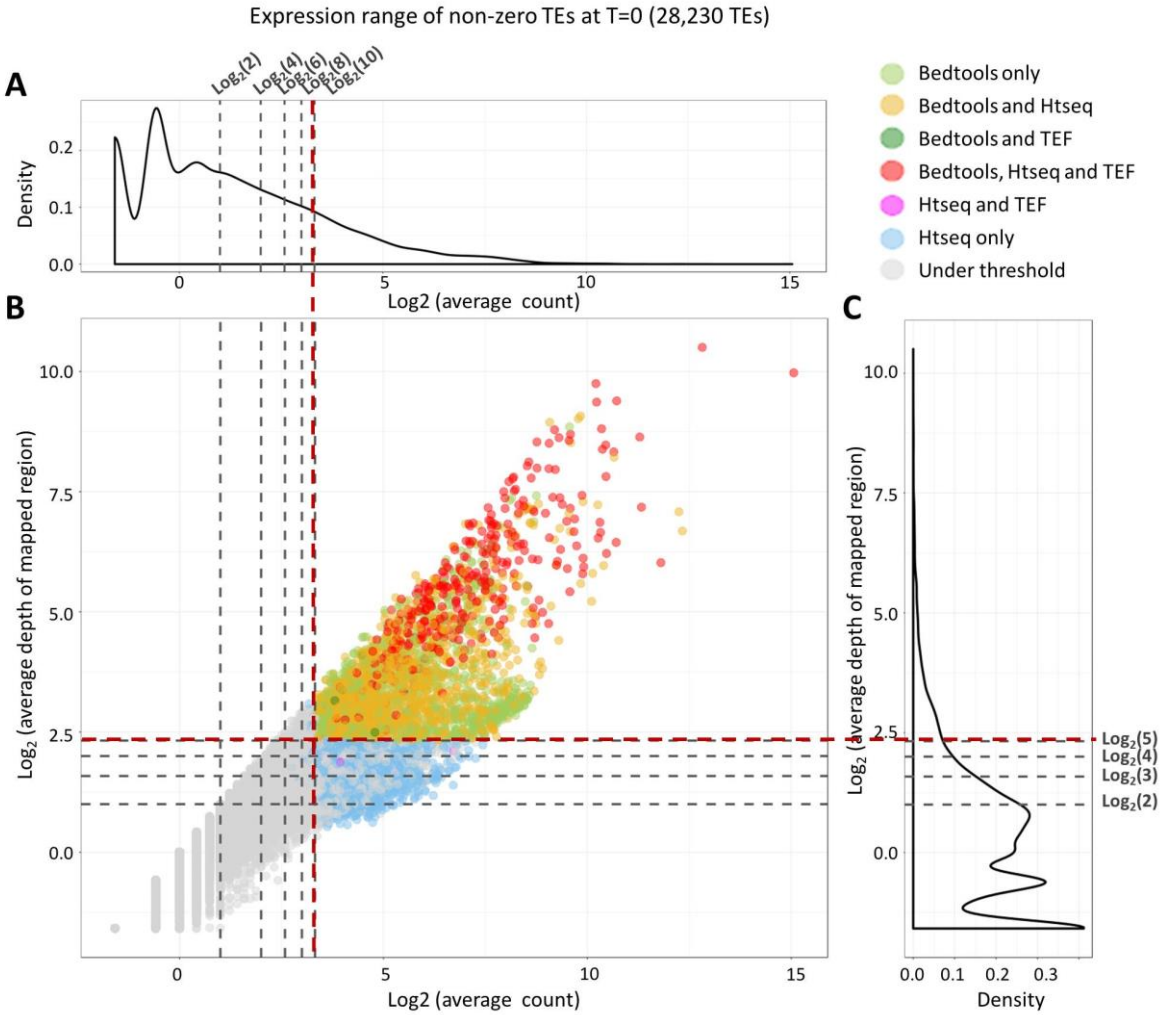



**Figure 2.8 Expression range of non-zero TEs at T=0**

Non-zero TEs include under-threshold TEs and expression candidates. The expression level of each individual was shown by plotting the logarithmically transformed read count against the logarithmic transformed average depth of the mapped region (centre). The density distributions of read count and average read depth were projected on the top and the right side, respectively. The cut-off thresholds in BEDtools sub-pipeline were indicated by red dashed lines, while grey dashed lines were added to indicate the position of other values lower than the thresholds. Expression candidates were coloured by the approaches where they were included. Under-threshold TEs were coloured in light grey.

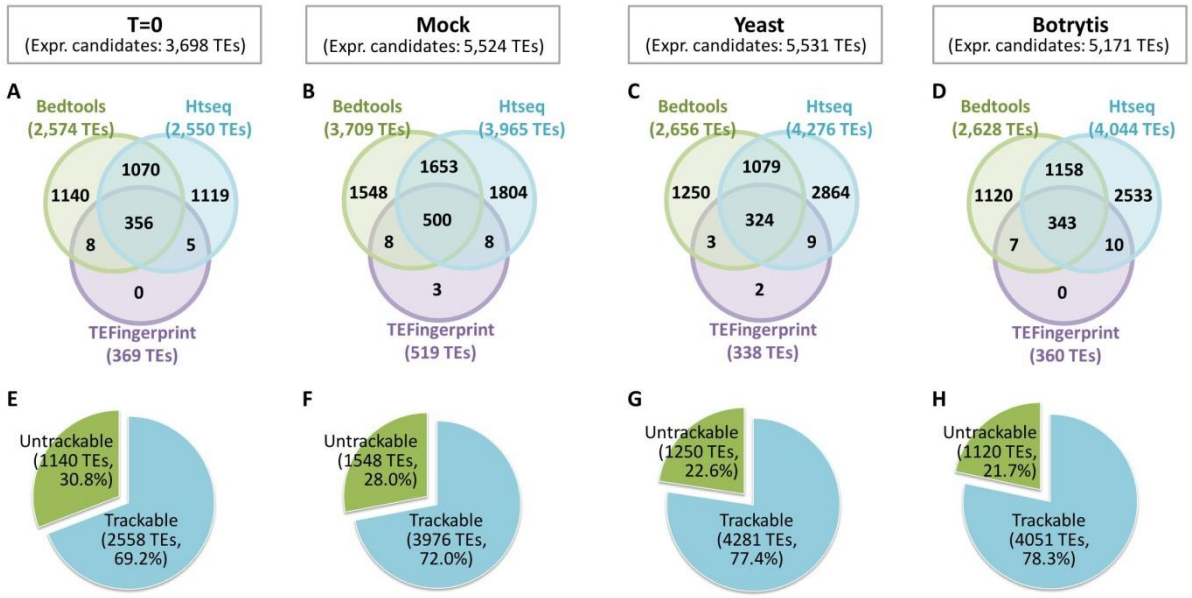



**Figure 2.9 Comparison of expression candidates collected by different sub-pipelines**

**(A-D)** Relations among three sub-collections of expression candidates obtained from the pipeline. The number of expression candidates included in each sub-pipeline was indicated. The overlapping areas denote TEs captured by multiple approaches. **(E-H)** Categorization of expression candidates by the presence of unique-mapping reads. Expression candidates with unique-mapping reads were trackable (blue), and the remaining candidates were untrackable (green).

Each set of expression candidates is the union of potentially expressed TEs captured by the three parts of the pipeline (Figure 2.9A-D). As BEDtools sub-pipeline captured TE loci containing unique- or multi-mapping reads over the given threshold (see 2.3 Methods), Htseq (`htseq-count`) sub-pipeline only identified TE loci that are uniquely mappable with over 10 unique-mapping reads (or 5 unique-mapping read pairs). Hence, as indicated by the overlapping areas of BEDtools and Htseq sub-pipelines in Figure 2.9A-D, the latter approach could also recognize a subset of expression candidates captured by BEDtools. The TEFingerprint sub-pipeline collects TE loci showing evidence of transcription across boundaries (Figure 2.3), as well as transcription of TE loci that are unannotated in our samples compared to the reference resource. The intersections of candidate pools collected by BEDtool and TEFingerprint sub-pipelines (Figure 2.9A-D) show that TEFingerprint supported recognizing the identity of highly conserved TEs transcribed across annotated TE boundaries. These Venn diagrams further reveal the proportion of trackable and untrackable expression candidates (Figure 2.9D-H). The trackable expression candidates denote expression candidates obtained through Htseq and TEFingerprint approaches, whose expression pattern could be traced by unique-mapping reads. The remaining expression candidates only mapped by multi-mapping reads were untrackable.

### 2.4.3 Comparison of the pipeline and TEtranscripts

TEtranscripts has been commonly applied to analyze the transcriptional expression of TEs by equal-weighting multi-mapping reads to each of the corresponding mapped loci, following by an expectation-maximization algorithm for estimating the relative abundance of each TE transcript before the integration with the unique-read counts and a final summarization of the total relative abundance for each TE family (Jin et al., 2015). On the other hand, the new pipeline aims at collecting all possible origins, i.e. individual TE locus, of TE-related reads without resolving the ambiguous alignments. To compare the pipeline with the well-established software TEtranscripts, the alignment data of polyadenylated transcriptome against V. vinifera genome generated by HISAT2 was fed into TEtranscripts to quantify TE's transcriptional activity and collect expressed TE families with a count threshold of 10.

For the embryogenic callus cultures at T=0, the 3,698 expression candidates captured by the pipeline originate from 174 of 232 TE families (Figure 2.10 A), 177 TE families were found to be expressed using TEtranscripts (Figure 2.10 B). A comparison between the two sets of expressed TE families showed 167 TE families were found expressed by both methods, whereas 7 and 10 families were uniquely identified by the new pipeline and TEtranscripts, respectively (Figure 2.10 C). Because the read count calculated by bedtools coverage represents the unweighted alignment result and shows all the possible origins of sequencing reads, the expression range of each individual TE locus included in the ten families uniquely found by TEtranscripts was examined by extracting their BEDtools read count and the average depth of mapped regions. Among the 3,117 TE loci included in the ten families, 2,811 of them had no mapped read, and the rest 306 TE loci were mostly covered by less than ten reads with the average depth of mapped regions all lower than 5 (Figure 2.10 D). Although the sum of the read count from individual TE loci reached the threshold of 10, this sum was connected with TE loci showing a considerable low expression level below the threshold applied for the individual. Therefore the 306 TE loci were remained excluded from the pool of expression candidates. On the other hand, the 7 TE families uniquely found expressed by our pipeline were corresponding to 11 expression candidates that were excluded by TEtranscripts due to two situations regarding the annotated TE features and the co-localized genes (Figure 2.10 E). The first situation is that the reads mapped to a TE also mapped to a gene's exon adjacent to the TE. The second one is that the TE-related reads also mapped to an overlapping exon. As two expression candidates belong to the first situation, the second situation corresponds to the other nine expression candidates, of which seven belongs to domesticated TE families (Vinesleeper-2, MUGvine-1 and MUGvine-2) of grapevine (Benjak et al., 2008; Knip et al., 2012). Although both situations involve TE-related reads originated from genes, the presence of these reads in the transcriptome may have epigenetic effect

on the TEs or the co-localizing genes. Therefore these 11 TEs were retained in the pool of expression candidates.

The comparison of expressed TE family obtained from the pipeline and TEtranscripts was also performed for mock (Figure 2.11), yeast (Figure 2.12) and Botrytis (Figure 2.13) treated EC cultures. Concordant with what was observed at T=0, the collections found by two methods were overlapped significantly (Figure 2.11C, Figure 2.12C, Figure 2.13C). Those families uniquely found in TEtranscripts collection obtain few TE individuals with mapped reads at a level close to background noise (Figure 2.11D, Figure 2.12D, Figure 2.13D). These TE loci would not pass the cut-off threshold individually unless they were analysed at the family level. On the other hand, TE families uniquely found by the pipeline were mostly comprised of expression candidates overlapping with expressed genes. (Figure 2.11E, Figure 2.12E, Figure 2.13E). Reads aligned to the TE features overlapping with genes were predominantly assigned by TEtranscripts to the genes instead of the TE loci. Because the overlapped annotation of TEs and genes in these cases might implicate the exaptation of TE-derived sequences into functional genes, we retain this kind of TE loci in the expression candidate pool as this will help to elucidate the location tendency of transcriptionally permissive TE loci in relation to genes (more details in chapter 3). In addition, the exclusion of a small number of families by TEtranscripts approach in yeast and Botrytis treatments were likely due to a third situation, in which expression candidates of these TE families merely passed the threshold of the pipeline but failed the threshold set for TEtranscripts.

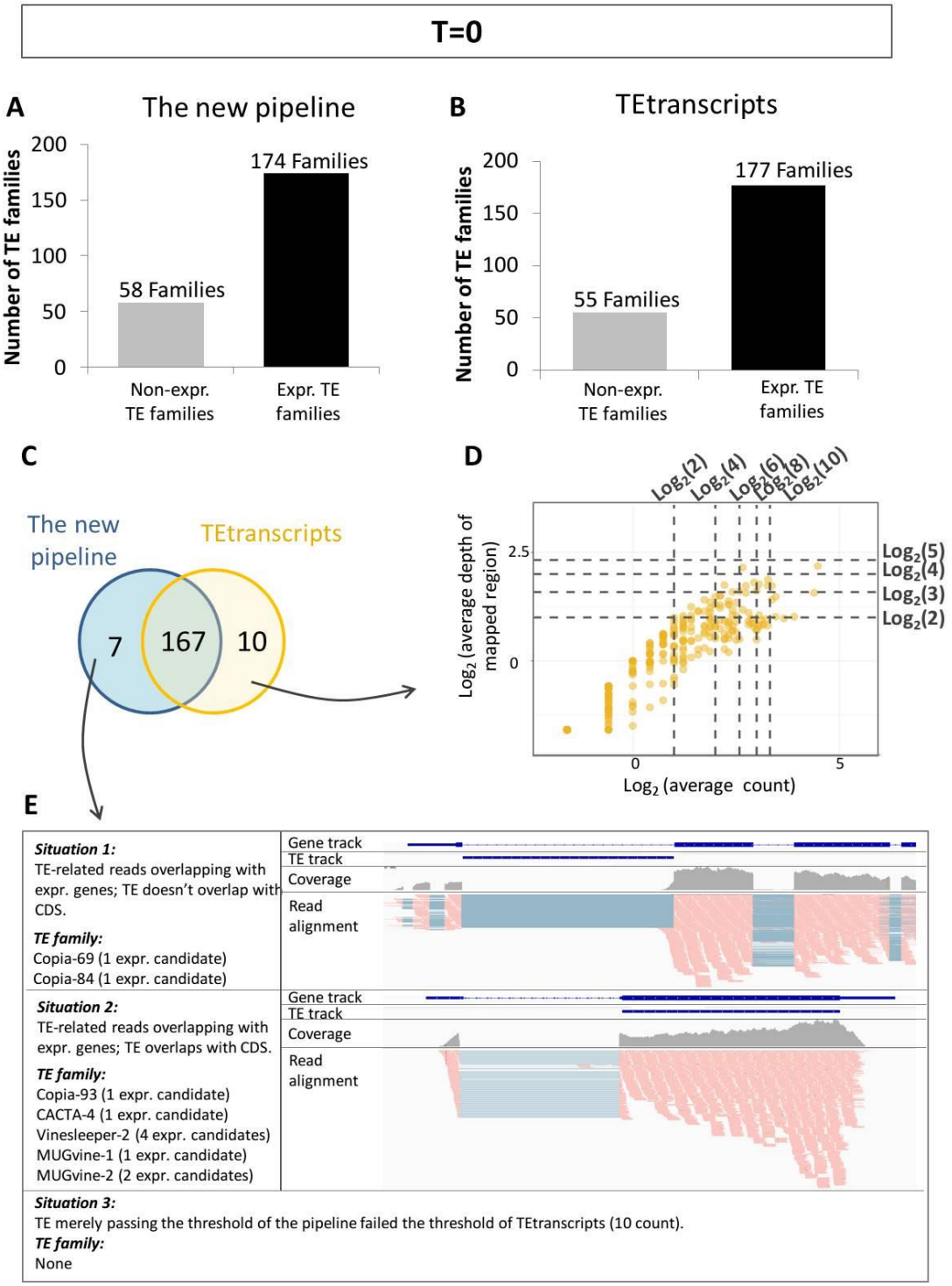



**Figure 2.10 Comparison of expr. TE families between the pipeline and TE transcrtipes.**

**(A)** With the new pipeline, 174 of 232 TE families were found obtaining expr. candidates. **(B)** Using TEtranscripts, 177 TE families were found active with an average read count higher than 10. **(C)** Comparison of active TE families captured by the pipeline and TEtranscripts. **(D)** The expression range of individual TE loci of the TE family uniquely found by TEtranscripts. **(E)** Reasons for the TE families uniquely included in the new pipeline were listed on the left and illustrated with some examples using IGV.

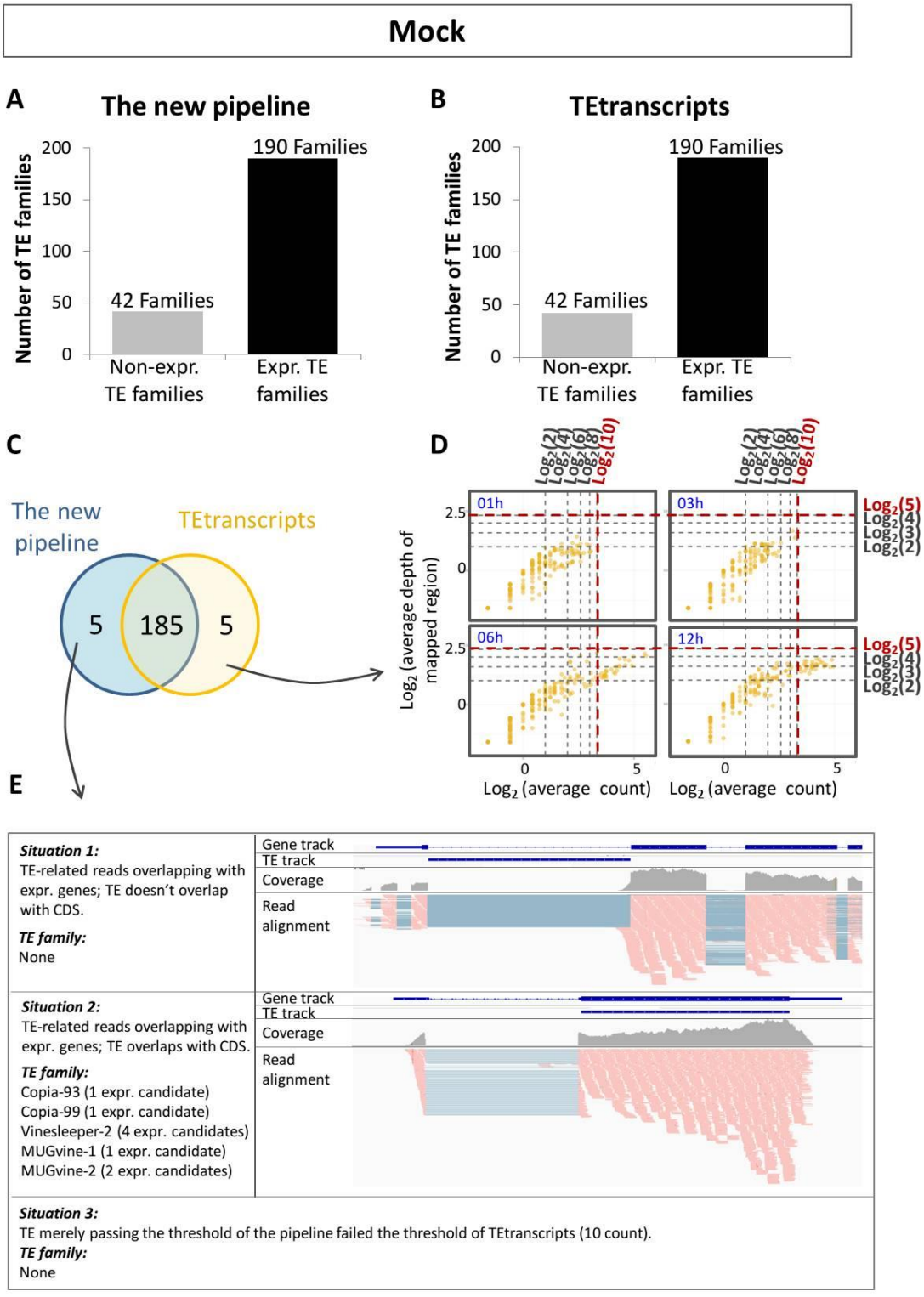



**Figure 2.11 Comparison of expr. TE families of mock treatment between the pipeline and TEtranscripts.**

**(A)-(B)** TE families were categorized by transcriptional activity identified by the pipeline (A) or TEtranscripts (B). **(C)** Comparison of active TE families captured by the pipeline and TEtranscripts. **(D)** The expression range of individual TE loci of the TE family uniquely found by TEtranscripts. **(E)** Reasons of the TE families uniquely included in the new pipeline.

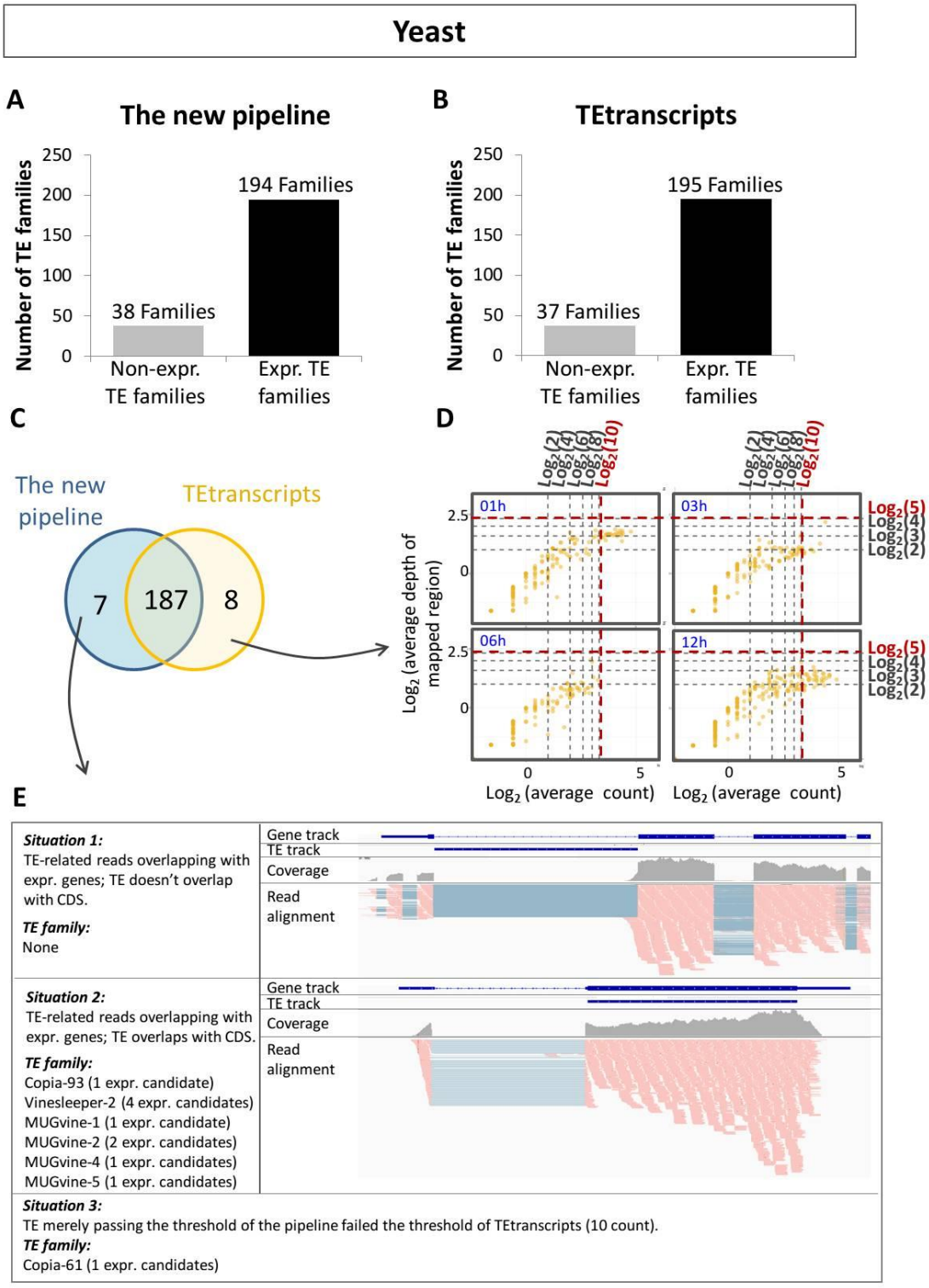



**Figure 2.12 Comparison of expr. TE families of yeast treatment between the pipeline and TEtranscripts.**

**(A)-(B)** TE families were categorized by transcriptional activity identified by the pipeline (A) or TEtranscripts (B). **(C)** Comparison of active TE families captured by the pipeline and TEtranscripts. **(D)** The expression range of individual TE loci of the TE family uniquely found by TEtranscripts. **(E)** Reasons of the TE families uniquely included in the new pipeline.

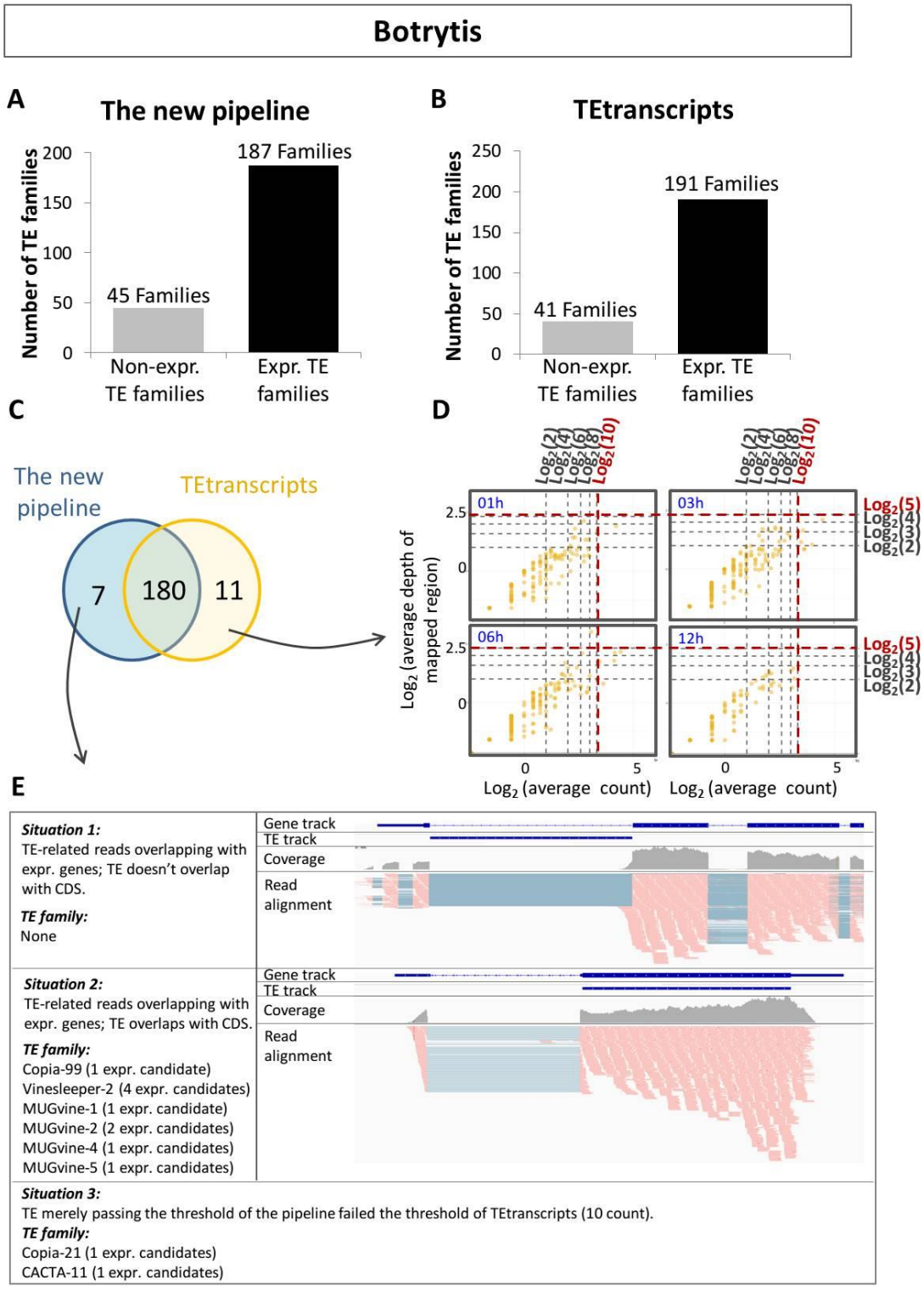



**Figure 2.13 Comparison of expr. TE families of Botrytis treatment between the pipeline and TEtranscripts.**

**(A)-(B)** TE families were categorized by transcriptional activity identified by the pipeline (A) or TEtranscripts (B). **(C)** Comparison of active TE families captured by the pipeline and TEtranscripts. **(D)** The expression range of individual TE loci of the TE family uniquely found by TEtranscripts. **(E)** Reasons of the TE families uniquely included in the new pipeline.

## 2.5 Discussion

### 2.5.1 Stress treatments

Following Lizamore's method (2013), before the application of live yeast or Botrytis cell extracts, solid EC samples were transferred to the liquid medium and shaken vigorously for 2 seconds to break the clusters of callus apart (also see 2.3.5). EC samples that only experienced vigorous shaking were collected as a control measure. However, the shaking procedure may have posed stress to EC as the separation of lumps of linked cells can damage cell wall integrity, therefore resembling wound treatment. It also means that EC treated with live yeast or *Botrytis* extract were, in fact, underwent a wound treatment and then continuous exposure to biotic stresses. For convenience, the experiment in which EC was only treated with wound procedure would be denoted as mock, while the other two with additional biotic stimuli are denoted as yeast or Botrytis treatments. The EC collected directly from the $C_1^P$ medium prior to any treatment would be represented as T=0.

### 2.5.2 The small proportion of TE-related reads

Following standard processing and alignment approaches for analysing RNA-seq data, the total reads mapping to the reference genomes range from 18 million to 40 million reads with a median of 36 million mapped reads (Table 2.2). Providing that the size of the sequencing reads is roughly 120 bp after adapter removal, this sequencing data can cover about 8X of the 500 Mb grapevine genome. In contrast to the substantial proportion of genomic sequences occupied by TEs (Table 2.1), only 0.5% to 2.5% of the mapped reads associated with annotated TE sequences (Figure 2.4). In rice, roughly 9% of the sequence-tagged connectors (STCs) were potentially derived from TEs (Mao et al., 2000). Even with 85% of the genome associated with TEs, maize only showed about 1.5% of the expressed sequencing tags (ESTs) associated with TEs (Vicient, 2010). It seems to be common that TE's contribution to the transcriptome is disproportional to the genomic sequences occupied by them.

### 2.5.3 Antisense reads of TEs in the transcriptome

For sequencing reads mapping to TEs, about 40% of them mapped in antisense orientation, indicating a substantial amount of TE-related antisense transcripts. These transcripts could be simply the read-through product transcribed from neighbouring genes or TEs with orientation opposite to the TE having antisense reads, such as the examples shown in Figure 2.5. Alternatively, the antisense transcription could be initiated from the antisense promoter within a TE, producing transcripts antisense to itself (Slotkin and Martienssen, 2007). The antisense transcripts can pair with the sense counterparts and form dsRNA, which could take parts in the biogenesis of natural antisense siRNAs (Borges and Martienssen, 2015). In addition to fueling the epigenetic suppression acting on TEs, it has been reported that the antisense promoter in a TE can initiate the antisense TE transcription that

extended outward to neighbouring genes, serving as their primary promoters (Nigumann et al., 2002) or participate in the epigenetic control of gene expression (Yang and Kazazian, 2006). Although these antisense TE transcripts can be crucial for TE and gene regulation, this research focuses on the sense transcripts that are more directly associated with TE's transcriptional activity. Therefore, only sense reads mapping to TEs were included in the analysis pipeline for collecting potentially expressed TEs (expression candidates).

### 2.5.4 Effective reduction of the search field for identification of active TE loci

Across all experimental conditions, the analysis pipeline has effectively excluded 75% to 87% of annotated TEs as having no mapping reads at all (Figure 2.6). Another 11% to 22% of annotated TEs were denoted as 'under threshold' and removed from the expression pool due to their inadequacy of read count or average read depth normalized by mapped region. The resulting 3600 to 5500 expression candidates accounted for 70% to 85% of the sequencing reads mapping to TEs (Figure 2.7). Although there are 15% to 30% of the TE-mapped reads associated with "under threshold" TEs, these reads scattered through 24000 to almost 50000 TE loci, leaving each of these mapped TE loci sparse read count or depth. As shown in Figure 2.8, the expression level of these "under threshold" loci was peaked at 0-2 mapped reads (Figure 2.8A) or 0-2 average read depth of mapped region (Figure 2.8C). With such a low expression level, the "under threshold" loci were very unlikely to be considered as transcriptionally active TE loci. In contrast, the majority of the TE-mapped reads were concentrated on a few thousands of expression candidates, each of which showed expression range clearly above the thresholds set in the analysis pipeline. Notably, some T=0 expression candidates are on the lower-right part of Figure 2.8. They display average read depth lower than 5 yet captured by Htseq sub-pipeline for each with 11 to roughly 180 unique-mapping reads at T=0. Therefore, this kind of TEs remained included in the collection of expression candidates.

### 2.5.5 Meaning of trackable and untrackable loci

In Figure 2.9, those TEs captured by Htseq or TEFingerprint sub-pipelines were binned into the 'trackable' group since the presence of unique-mapping reads confirmed the transcriptional activity derived from them, and their expression dynamics can be indicated by these unique-mapping reads. Despite that the unique-mapping reads may only cover part of the TE features with unique polymorphisms, the dynamic read depth of this unique part is supposed to generally reflect the overall TE expression change in response to different experimental condition. Therefore, the differential expression change of these TEs can be tested statistically.

By contrast, expression candidates lack of unique-mapping reads were assigned as 'untrackable', providing that a TE locus in this group is one of the multiple "suggested" origins of those multi-

mapping reads aligned to it. Without unique-mapping reads as confirmation and indicator of expression, there is no way to track the expression changes of these individual loci. However, this is the category that is likely to contain those active TE insertions that are evolutionarily newly generated or competent for autonomous transposition. This kind of TE loci derived from the same TE family are assumed to be highly similar or even identical. The transcriptional activation from one or some of them would lead to ambiguous alignment, resulting in multiple associated TE loci being categorized into the un-trackable group. In other words, the untrackable group may reveal the transposition activity of the competent TE family.

### 2.5.6 Combined usage of the analysis pipeline and TEtranscripts

The high consistency of the expressed TE families between the analysis pipeline and TEtranscripts implicates the possibility to explore TE behaviour with the help of these two methods. For example, while TEtranscripts can efficiently quantify TE expression at the family level, which can be further applied for differential expression test, the analysis pipeline can pinpoint the possible origins in the reference genome that contribute transcripts associated with the specific TE family. With additional characterization steps, some of these possible origins can be used to analyse the dynamic and differential transcriptional activity and facilitate the investigation of the transcriptional association between TEs and genes. Alternatively, among up-regulated TE families identified by TEtranscripts, the analysis pipeline can help to narrow down the searching scale of potentially autonomous loci that are likely to produce functional transcripts necessary for mobilization. For instance, TE activation has long been investigated at the family level in *Arabidopsis* impaired in epigenetic silencing (Cavrak et al., 2014; Lanciano and Cristofari, 2020; Yu et al., 2017), and TE families contributing to a TE storm in the *Drosophila* model of amyotrophic lateral sclerosis (ALS) has been characterized (Krug et al., 2017). Identification of the individual active TE loci in these models can shed light on the impact of TE activation on neighbouring genes or the pathogenesis regarding the balance between gene and TE expression. In chapter 5, the ability of this pipeline in recognizing individual active TE loci were explored in *Arabidopsis* and *Drosophila* datasets that have shown transcriptional activation of TEs.

## 2.6 Conclusions

The purpose of this analysis pipeline is neither resolving the ambiguous alignment of TEs nor quantifying the exact expression level of each TE locus. Instead, it focuses on capturing potentially expressed TE loci as much as possible from multiple perspectives. It has shown that it is possible to distinguish a group of TEs with transcriptional potentials from the large pool of annotated repetitive elements. It cannot provide the overall expression level of a TE having multi-mapping reads, yet it allows investigation of a specific experimental treatment on TE's transcriptional activity if this TE is also included by the Htseq or TEFingerprint sub-pipelines. This pipeline also facilitates the search for TE loci that may generate autonomous transcripts for mobilization. Overall, this pipeline can increase the granularity of RNA-seq analysis for TE-related study.

# Chapter 3

# Characterization of potentially expressed transposable elements

## 3.1 Overview

For a long time, activation of transposable elements (TEs) has been linked with "selfish" proliferation and sabotage of gene function. However, the mutagenic ability and stress responsiveness of endogenous TEs can be leveraged to enrich polymorphisms of crop populations and, thus, facilitate crop improvement. In Lizamore's research (2013), transcriptional activation of endogenous TEs has been observed in *Vitis vinifera* embryogenic callus inoculated with extracts of *Botrytis cinerea* or live *Hanseniaspora uvarum* cultures. Although TE families that showed an increase in transcript levels in this study have been identified, it is unclear which of the individual TE loci of these families contributed to the transcriptional activation. Identification and characterization of individual active TE loci can reveal factors that are important to boost mobilization efficiency.

With this aim, this chapter focuses on recognizing factors crucial to TE's transcriptional activation by using the experimental system pre-developed on the *P. noir* clone UCD5, as well as the analysing pipeline established in chapter 3 that have identified a group of potentially expressed TE loci (hereinafter 'expression candidates'). Two LTR-TE families, Copia-3 and Copia-23, were found to sustain the high integrity and identity in the annotated full-length loci and show the transcriptional potential of autonomous mobilization. The location distribution of all expression candidates reveals a general trend of TE activation favouring TE loci in genic regions, particularly in introns of expressed genes. This trend is also observed in full-length Copia-3 and Copia-23 loci potentially contributing autonomous transcripts. These findings suggest that the intragenic area with transcriptionally relaxed status might benefit TEs within it by offering transcriptional opportunity and likely pose as a shelter for intact TEs from complete suppression. The TEs that take advantage of the intragenic area for their own transcriptional activation might consequentially tend to re-insert into areas of similar transcriptional status due to the structural accessibility of these regions.

In addition, the clipped-read-based approach reveals the possibility of TE activation from unannotated TE loci and the discrepancy between *P. noir* and the reference genome. This also emphasizes the need for long-read sequencing technology to resolve genomic areas enriched with repetitive sequences particularly.

## 3.2 Introduction

### 3.2.1 Why is it important to investigate the landscape, properties and prerequisites of endogenous TE transcription?

Transposable elements contribute a substantial proportion of eukaryotic genomes. The human genome sequencing project launched in 1990 revealed that repetitive sequences occupy nearly 50% of the genome (International Human Genome Sequencing Consortium, 2001), while 37.5% of the mouse genome is considered transposon-derived (Mouse Genome Sequencing Consortium, 2002). In plants, TE-related fraction can range from 90% in the barley genome to as little as 25% in *Arabidopsis* (Tenaillon et al., 2010). These elements have for a long time been described as parasitic DNA and frequently linked with diseases such as cancer and neuronal disorders, as well as aberrant gene expression (Chuong et al., 2016; Feschotte, 2008; McConnell et al., 2017). However, there has been a recent resurgence in interest in the roles of TEs, in particular associated with genome and organism evolution (Casacuberta and González, 2013; Feschotte, 2008; Lisch, 2013; Tenaillon et al., 2010; Zhao et al., 2016). In the evolutionary process, often deleterious TE insertions are removed from host genomes by purifying selection, whereas some TE insertions benefit the host cell through changing expression patterns and properties of co-localized genes (Chuong et al., 2017). Chapter 1 has described various studies depicting the exapted roles of TEs in providing *cis*-regulatory elements at the promoter region of co-localized genes that consequently confer stress-responsiveness and offering functional protein domains domesticated into DNA-binding protein or transcriptional factors (Feschotte, 2008). In addition to the genetic effect of TE mobilization, dynamic transcriptional activity and altered distribution of TEs in the host genome can affect the landscape of epigenetic hallmarks acting against TEs (see section 1.3), and therefore contribute to epigenetic plasticity, especially for TE insertions within introns or promoter regions of genes (see section 1.4). The nature of endogenous TEs to generate genetic and epigenetic polymorphisms has been considered a powerful tool to accumulate phenotypic variation. To harness this tool, TE's stress-responsiveness, conveyed through the *cis*-regulatory elements harboured within TE sequences, have been found as leverage to stimulate TE transcription (see sections 1.4.2 and 1.5), which is the necessary step to initiate the mobilization of both type I and type II TEs (see section 2.2.2.2).

As introduced in section 1.5, re-activation of TE transcription and mobilization by tissue culture as well as biotic or abiotic stressors has been recorded in various plant species, including grapevine (*Vitis vinifera*). As a long-term goal of crop improvement, the mutagenic and stress-responsive nature of endogenous TEs was leveraged by Lizamore (2013) to establish new clonal material of grapevine (see section 1.1). In Lizamore's research (2013), a number of endogenous grapevine TE families in embryogenic callus were found to be responsive to a range of biotic stressors, in particular cell

extracts of *Botrytis cinerea* and live cultures of *Hanseniaspora uvarum*. The transcriptional activation of four families of LTR retrotransposons (LTR-TEs), including *Cremant* (*Copia-30*), *Edel* (*Copia-31*), *Noble* (*Copia-3*) and a Gypsy element *Gret1*, were detected by real-time RT-PCR. New insertions of these four TE families were discovered in regenerated plants by the PCR-based methodology, sequence-specific amplification polymorphism (S-SAP). These findings reveal the potential of using endogenous TE and tissue culture systems as a bvvvrfrrdeszhayZpathway for crop improvement. However, it is estimated that the mobilization rate of these 4 LTR-TE families is fewer than one new TE insertion event per genome (Lizamore, 2013).

In order to understand how to increase the mobilization efficiency of endogenous TEs, it is important to be cognizant of the prerequisites for TE activation at the granularity of individual loci. As previously mentioned, the first step for both type I and type II autonomous TEs to initiate TE mobilization is the transcription of TE DNA into mRNA, which is further required for the synthesis of TE proteins mediating mobilization (see sections 1.2 and 2.2.2.2). However, identification of transcriptionally active TE loci has been encumbered by the repetitive and self-proliferated nature of TEs, and thus most of the analysis tools for TE transcription study are only applicable for the identification of a group of TE loci or applicable at the TE family level (section 2.2.2.2). While these tools individually collect a subset of transcriptionally active TE loci, an analysis pipeline that chains up the existing methodologies have been established to maximize the identification of expressed TE loci (Figure 2.2). This pipeline has been tested on the polyadenylated transcriptome data (RNAseq data) derived from grapevine embryogenic callus subjected to time-series stress treatments (Figure 2.1), which include a wound-like pre-treatment to increase the accessibly of cells to the biotic stressors and an inoculation either with *B. cinerea* cell extracts or *Hanseniaspora uvarum* live cultures. Depending on the treatments, the analysis pipeline has successfully identified 3,000 to 5,500 potentially expressed TE loci (hereinafter expression candidates) from the total 223,411 TE loci annotated in the *V. vinifera* reference genome (Figure 2.6). The characteristics of these expression candidates are thoroughly investigated in this chapter with the purpose to comprehend the prerequisites for individual TEs to be transcriptionally active.

## 3.2.2 Factors that contribute to TE mobility

To explore the characteristics of transcriptionally active TEs, expression candidates collected from the pipeline described in chapter 2 were surveyed from the followings perspective:

### 3.2.2.1 Element Size

A burst of activity of a transposon family can lead to hundreds or thousands of identical insertions throughout the host genome. However, as elements age in the host genome, polymorphisms

increase in the form of single nucleotide variances (SNVs) as well as insertions and deletions (INDELs), which can gradually cripple TE's competence in terms of the ability for autonomous mobilization. Among these forms of mutations, the erosion of element size by deletion event is the most straightforward indication of TE's (inability for) mobility. Therefore, in this chapter, one of the method to determine if an element at a particular locus may be autonomous or not is to examine the size integrity of any annotated element in the genome by comparing its length to the canonical sequence for that element family. For instance, if a TE locus is 10% shorter than the length of its corresponding canonical element, it is less likely to retain its intact functionality.

#### 3.2.2.2 Location

Although TEs are frequently targeted by the host's epigenetic silencing systems and transcriptionally suppressed, TEs inserted within promoters, exons and introns of genes, as well as intergenic region distal to genes, might exhibit different degree of transcriptional restriction caused by different levels of epigenetic silencing against these TEs. This is due to the dual need for restriction of TE transcription without compromising gene expression (see sections 1.4 and 1.6.2). For intronic TEs, Saze et al. (2008) found that the epigenetic marks deposited on these TEs are important to prevent aberrant alternative splicing and premature transcriptional termination of host genes. However, Le et al. (2015) found that genes that house TEs in introns had lower expression levels than genes without TEs, and the expression level of these host genes is inversely related to the DNA methylation level of the intronic TEs, suggesting that the epigenetic suppression on these intronic TEs might prevent the host genes from being expressed at a high level. In some of the cases of TE insertions in promoter regions, TEs serve as *cis*-regulatory elements and lead to exaptation of stress-responsiveness for co-localized genes (see sections 1.4.2 and 1.5). De-repression of these TEs at promoter regions would be critical for the transcriptional activation of nearby genes. Taken all together, while deep silencing of intragenic TEs is possible to cost cells important host genes co-repressed with the TEs, host cells might tune the level of silencing acting on intragenic TEs to retain the basal function of genes with permitting TE expression. In other words, transcriptionally active TEs might predominantly co-localize with expressed genes.

#### 3.2.2.3 Structural components

The structures of different types of TEs have been described in chapter 1. Given that each of these structural components has a specific function to facilitate transposition, the presence, absence or mutation of these can allow the determination of the competency (or not) of any given element. Given the variability of these structural components among TE families, this part of the analysis will focus only on the predominant types of TEs (LTR and non-LTR retrotransposons as well as TIR-type

DNA transposons) and their structural components commonly accepted to allow identification of putatively autonomous TE loci.

***LTR retrotransposons***

This type of TE is characterised by the presence of flanking identical long terminal repeat (LTR; see Figure 1.1 A). As elements age at any given locus, the LTR pair independently and gradually accumulate mutations. As a result, the presence of an identical LTR sequence has been used as an indicator for very recent mobilization and, therefore, competency of the element to transpose. Correspondingly the level of sequence divergence can be used to estimate the age of an element (Schulman, 2013). The more diverse the two LTR sequences, the less recent the insertion is. In general, the *de novo* transcription of an autonomous LTR-TEs initiates from the transcription start site (TSS) within 5' LTR (Figure 1.1 A), covering the internal domain (INT), and stops at the 3' LTR. When analyzing RNA-seq data, in the safe end to include potentially active autonomous LTR-TEs, those with >90% of INT covered by short RNA sequencing reads would be collected as potential origins producing competent transcripts (Figure 3.1).

***Non-LTR retrotransposons***

LINE elements are the most prominent super-family in this category. Most autonomous LINEs consist of two open reading frames (ORF), one of which encodes the reverse transcriptase (RT) with RNAase activity (Schulman, 2013). Therefore the presence of ORF encoding RT in the canonical sequence of a LINE family is a criterion of autonomous mobilization. The transcription initiates from the beginning of the 5' untranslated region (UTR) and terminates at the poly-A sequence attached downstream of the 3' UTR (Figure 1.1 B). To identify putative autonomous LINE containing loci with the potential of full transcription from RNAseq result, the breadth of coverage of the entire full-length LINE should reach 90% (Figure 3.2).

***Terminal Inverted Repeat (TIR) -DNA transposon***

This class of TEs (hereinafter TIR-TEs) is the most dominant category in type II DNA transposon. They are known for the presence of terminal inverted repeats (TIRs) flanking the TE feature (Figure 1.1 C). Internally there should be an ORF that encodes an enzyme known as transposase (TPase), which recognizes the TIR sequences and executes the excision and reintegration of the element (Wicker et al., 2007; also see section 1.2.2). Therefore, an autonomous TIR-TE locus should retain TIRs and the TPase-encoding ORF (Figure 3.3). The transcription and translation of this ORF are required to achieve self-competent mobilization of an autonomous TIR-TE. In order to utilize the aforementioned RNAseq data to identify autonomous TIR-TE loci that are likely to achieve mobilization, it is assumed that these loci would obtain >90% breadth of coverage across the ORF encoding TPase (Figure 3.3).

## 3.3 Methods

### 3.3.1 TE integrity analysis

The length of each annotated TE locus was compared to the length of the corresponding canonical element sequence retrieved from the Repbase update database and reconstructed as described in chapter 3.3.3. TE loci longer than 90% of the corresponding canonical elements in length were considered full-length elements, and the rest of the annotated loci were grouped as fragmented TEs.

### 3.3.2 Identification of transcriptionally active TE family

To generate figures like Figure 3.5 D, expression candidates were initially grouped into trackable (collected from the sub-pipeline 1 and 3 in Figure 2.2) and un-trackable (captured by the sub-pipeline 2 in Figure 2.2 only) expression candidates. All expression candidates were further binned by family and integrity. This information was then integrated into a chart like Figure 3.5 D, where, for each TE family (the y-axis), the accumulated number of un-trackable and trackable expression candidates are respectively plotted on the left and right sides of the chart; based on this layout, the accumulated number of full-length and fragmented loci is indicated by dark and light colours respectively.

### 3.3.3 Cladogram analysis of full-length Copia-3 and Copia-23

In the previous analysis (section 3.3.2), Copia-3 and Copia-23 were found to be the two TE families representing the most un-trackable full-length expression candidates (Figure 3.5 – Figure 3.8), suggesting that these two TE families retain a substantial number of highly similar or identical full-length loci that are indicative of their recent activity in the evolutionary time. To investigate how diverse are these full-length TE loci, multiple alignments of the canonical sequences and full-length TE loci of Copia-3 and Copia-23 was performed by Geneious using MUSCLE alignment option with default settings. The Neighbour-Joining consensus tree was then constructed by Geneious with 100 bootstraps and a 90% support threshold. The tree was further illustrated by iTOL (Interactive Tree of Life, https://itol.embl.de/). All the analysed sequence here are labelled by five different colours (Figure 3.9 and Figure 3.10) that represent five categories of these full-length elements. These categories are as follows:

(1) structurally autonomous (full-length and flanked by LTRs) untrackable expression candidates;

(2) structurally autonomous trackable expression candidates;

(3) structurally non-autonomous (full-length but lost at least one LTR) un-trackable expression candidates;

(4) structurally non-autonomous trackable expression candidates;

(5) non-expressed full-length loci.

Note that in this analysis, LTR-TE loci that are full-length and flanked by LTRs are denoted as 'structurally autonomous' loci, while LTR-TE loci that are full-length but not flanked by LTRs are denoted as 'structurally non-autonomous' loci for convenience. Although there are other sequence compartments that are also determinant to the autonomous mobilization of LTR-TEs, such as primer-binding sites as well as *gag* and *pol* genes encoding proteins required for self-competent transposition, the selection of LTR-TE loci that are full-length in sequence size and flanked by LTRs has efficiently excluded 78.4% and 72.6% of the total annotated Copia-3 and Copia-23 loci, respectively, leaving 87 Copia-3 loci and 177 Coipia-23 loci considered as structurally autonomous loci. The polymorphisms that accumulated in sequence context but did not greatly affect sequence size were then measured during sequence alignment by MUSCLE, and we took this into consideration in the construction of the Neighbour-Joining consensus tree.

### 3.3.4 Analysis of reads mapping to Copia-3 and Copia-23

To test whether reads mapping to Copia-3 and Copia-23 expression candidates can possibly all derived from fragmented un-trackable expression candidates instead of full-length un-trackable expression candidates that share similarity with the fragmented counterparts, these reads were collected using bedtools-intersect (Quinlan and Hall, 2010) and analysed as follows: reads from the triplicates of same time point (see the experimental design in section 2.3.1) were merged and then categorized by their mapping destinations:

(1) fragmented un-trackable expression candidates;

(2) full-length un-trackable expression candidates;

(3) fragmented trackable expression candidates;

(4) full-length trackable expression candidates.

This gave four groups of reads for plotting Venn diagrams using the R package VennDiagram (Chen and Boutros, 2011). Note that, as presented in the Venn diagrams in Appendix C.4, a multi-mapping read may be categorised into multiple categories, including the categories of trackable expression candidates for that part of the DNA sequences of these trackable expression candidates may be identical to a portion of the DNA sequences of the un-trackable expression candidates.

### 3.3.5 LTR domain annotation

To identify full-length LTR-TE loci retaining LTRs and estimate the insertion date of these loci (the following section), the coordinates of LTRs in the *V. vinifera* reference genome were extracted based on the LTR domain annotation that has been established by Lizamore (2013). Without reassembling of LTR-INT-LTR structure, TE element sequences of *V. vinifera* retrieved from the Repbase database were adopted directly for the LTR annotation following workflow described in Lizamore (2013) and chapter 2.3.3. This kept annotated LTR and INT domains separated. The previous annotation version generated by using reconstructed LTR-INT-LTR canonical sequences was compared with this version (LTR and INT separated) of annotation to identify local copies of LTR-retrotransposons (LTR-TEs) flanked by LTRs on both ends and extract the coordinates of LTR domains.

### 3.3.6 Estimation of LTR-TE insertion date

Full-length TE loci flanked by LTRs were considered intact copies. The insertion time of these TE loci was dated by measuring the divergence between the 5' and 3' LTR for each locus, as proposed by SanMiguel et al. (1998). With the coordinates of INT domains in the reference genome, the sequences of paired LTRs were extracted using `bedtools getfasta` (Quinlan and Hall, 2010) and aligned by MUSCLE (Edgar, 2004) with the settings `-distance1 kmer4_6 -clwstrict`. Following Vitte et al. (2007), the observed divergence was corrected according to the Jukes-Cantor sequence evolution model (Jukes and Cantor, 1969) with the formula:

$$K = \frac{-3}{4} ln\left(1 - \frac{4}{3}p\right)$$

where K is the corrected divergence and p is the proportion of different bases in the two LTR sequences. The insertion date was then translated from the corrected divergence with an average substitution rate of $6.5 \times 10^{-9}$ substitution per site per year estimated from the *Adh1* and *Adh2* genes of grass species including maize, rice and barley (Gaut et al., 1996) and had been adopted by Moisy et al. (2008) for the grapevine genome. The distribution of insertion time was plotted by ggplot2 (Wickham, 2016) and depicted the insertion history of LTR-TE families with at least 10 intact copies. The peak mobilization was estimated from the distribution of insertion times using the R package hdrcde (Hyndman, 2018). To test whether the LTR-TE families, Copia-3 and Copia-23, stacked with un-trackable full-length candidates were more recently active than others, their peak insertion dates were compared with other 5 LTR-TE families with trackable full-length candidates and a statistical test (t-test) of the mean insertion times was performed in a pair-wise manner with the function t-test in R.

### 3.3.7 Location bias analysis

The location of annotated TEs in association with genes was analysed using `bedtools intersect`. The intersects between TEs and gene-related features, including exon, intron, 2kb upstream (N-flank) and 2kb downstream of any given gene locus, were further examined in R. If a TE overlapping with an exon and intron has over 95% of its body covered by the intron, it would be assigned as intronic TE, and otherwise, it would be assigned with TE loci overlapping with exon. This rule was also applied to a TE overlapping with an exon and flanking region. All annotated TEs and expression candidates were categorized with R package dplyr (Wickham et al., 2018) and plotted layer by layer as a pie graph using ggplot2 (Wickham, 2016). The goodness of fit X-square test was performed using the function `chisq.test` implemented in the R package MASS (Venables and Ripley, 2002).

### 3.3.8 Identification of potentially autonomous expression candidates

For type I LTR-retrotransposons (LTR-TEs), the transcription initiates from within the 5' LTR and progresses through the primer binding site (PBS), internal domain (INT) encoding proteins necessary for autonomous transposition, as well as polypurine tract (PPT), and termination at the 3' LTR (Figure 3.1 A). Therefore, as shown in Figure 3.1 B, the workflow for selecting putative autonomous LTR-TE candidates exhibiting potential competent transcription began with the selection of full-length TEs showing more than 90% of length coverage relative to the corresponding canonical TE sequence. Secondly, full-length candidates were examined for the presence of a pair of LTRs. Lastly, using `bedtools coverage` (Quinlan and Hall, 2010), only candidates with > 90% breadth of coverage over the INT domain would be qualified as autonomous expression candidates.

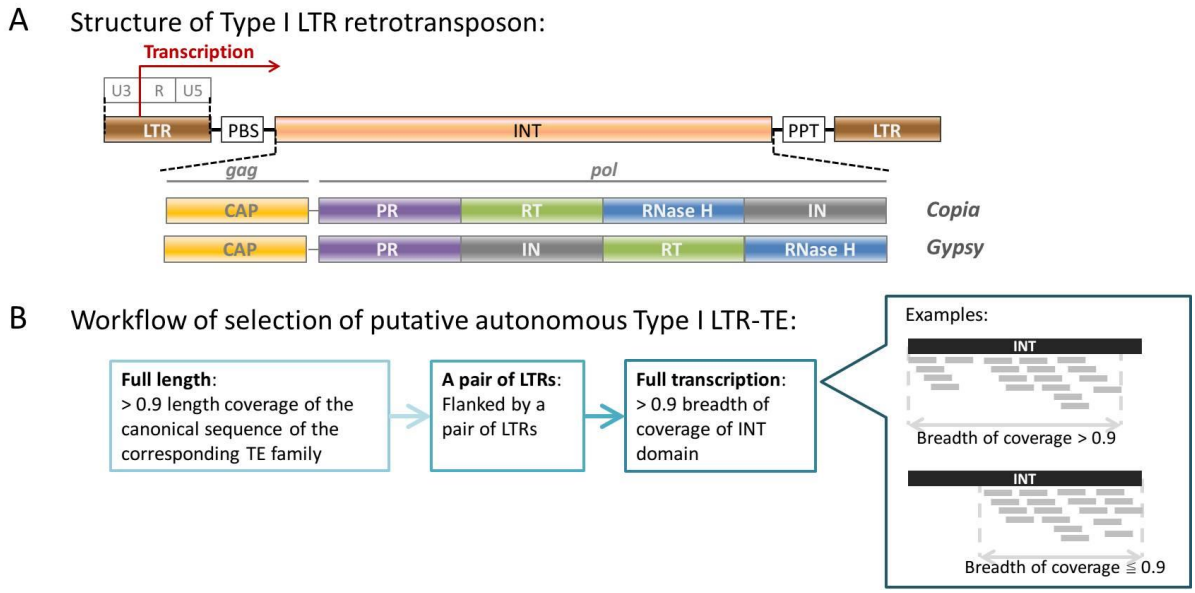



**Figure 3.1 Identification of autonomous LTR-TE expression candidates**

**(A)** Structure of LTR-TEs. The 5' LTR comprises unique 3' (U3), repeated (R) and unique 5' (U5) regions, which followed by primer binding site (PBS) for reverse transcription. The internal domain (INT) contained *gag* and *pol* genes encoding capsid-like-proteins (CAP), protease (PR), reverse transcriptase (RT) with RNase H activity, and integrase (IN). The 3' LTR is identical to the 5' LTR and sits after the polypurine tract (PPT). **(B)** Workflow for collecting potential origins of autonomous LTR-TE transcripts. The short grey segments denote sequencing reads.

Autonomous type I non-LTR retrotransposons (non-LTR-TEs) were mostly LINE elements possessing 5' and 3' untranslated regions (UTR) and an open reading frame (ORF) encoding poly-protein that can be processed into a protease (PR), reverse transcriptase (RT) and RNase H (Figure 3.2 A). The transcription initiating from the 5' UTR throughout the elements is necessary for reverse transcription and autonomous transposition. Therefore, full-length expression candidates of non-LTR-TEs were initially collected, following by a selection for those originated from non-LTR-TE families retaining intact RT domain with putative active sites. Those with over 90% breadth of coverage throughout the elements were considered autonomous expression candidates.

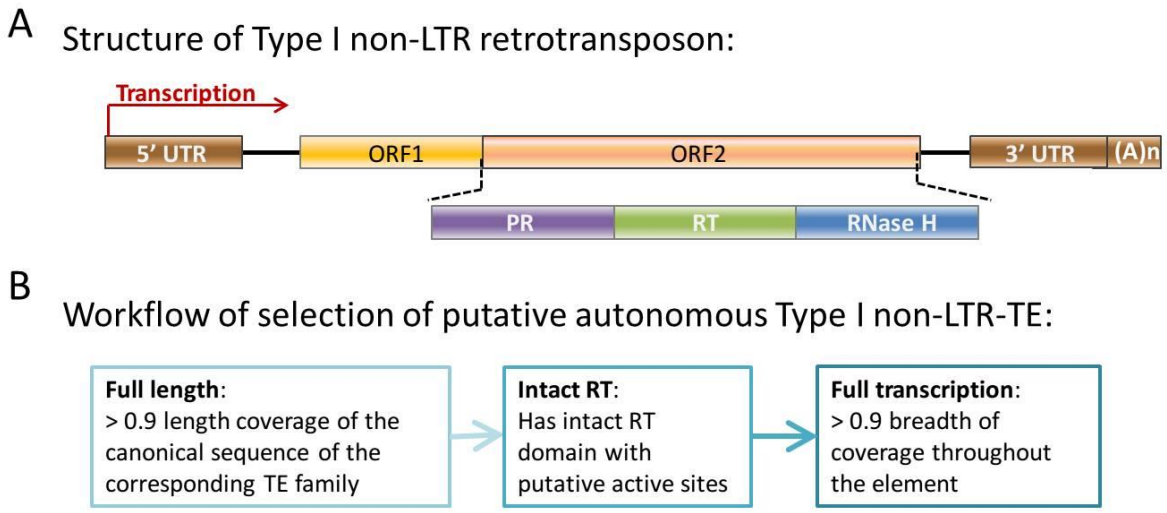



**Figure 3.2 Identification of autonomous non-LTR-TE expression candidates**

**(A)** Structure of non-LTR-TEs. Starting from the 5' UTR, non-LTR-TEs normally obtain two ORFs as the longer one encodes a poly-protein that can be cleaved into protease (PR) and reverse transcriptase (RT) with Rnase H activity. The polyadenylation signal ((A)n) is following behind the 3' UTR. **(B)** Workflow for collecting potential origins of autonomous LTR-TE transcripts.

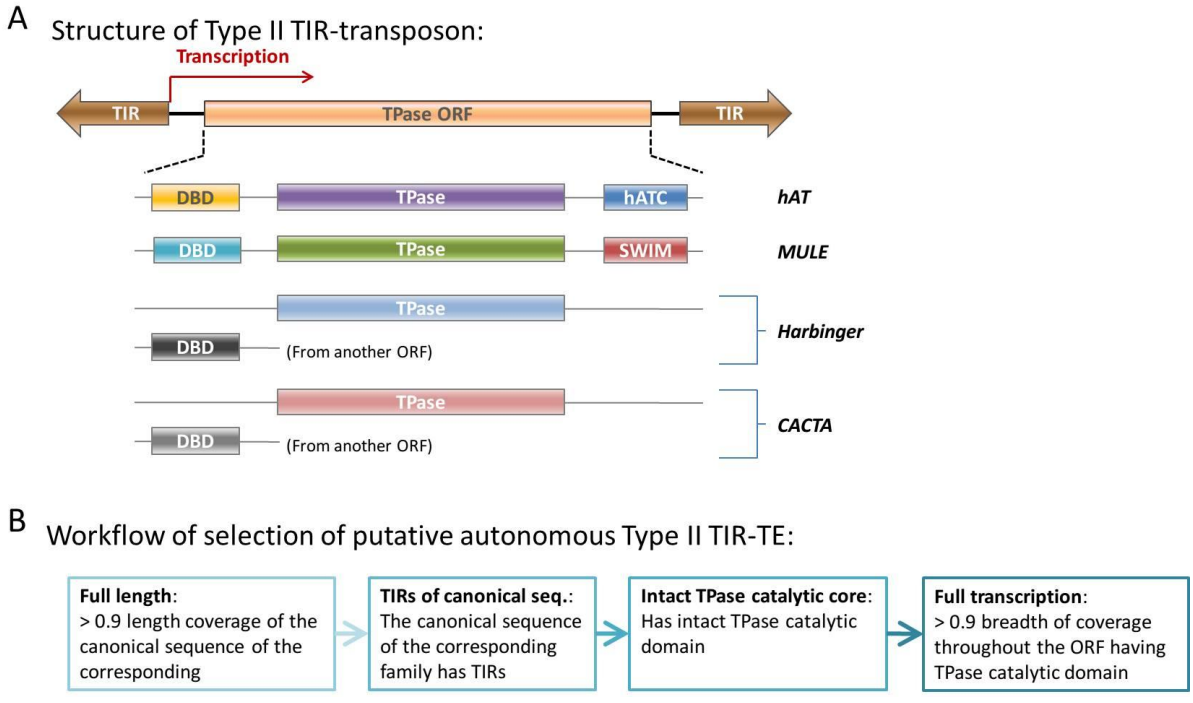



**Figure 3.3 Identification of autonomous TIR-TE expression candidates**

**(A)** Structure of TIR-TEs. Flanked by TIRs, the main structure of TIR-TEs is the ORF encoding transposase (TPase) composed by the TPase catalytic core and other functional domains depending on the families. **(B)** Workflow for collecting potential

origins of autonomous TIR-TE transcripts. DBD, DNA-binding domain; hATC, *hAT* C-terminal dimerization domain; SWIM, Zn-chelating domain of SWI2/SNF2 and MuDR.

Type II expression candidates are typically flanked by terminal inverted repeats (TIRs) and contain an ORF that encodes a transposase (TPase) (Figure 3.3). As previously shown in section 1.2.2, the TPase protein is assembled with multiple functional domains, including DNA binding and TPase catalytic cores, as well as other domain for metal ion chelating or dimerization depending on the TE families, such as hATC (*hAT* C-terminal dimerization) domain for *hAT* and SWIM (SWI2/SNF2 and *MuDR*) for *MULE* (Feschotte, 2008). For some of these TE families, however, the TIRs structure has been missing in the canonical element sequences, leaving them incompetent for autonomous mobilization. Therefore, only full-length elements belonging to the competent TIR-TE families were included in this study. The coordinates of the ORF encoding TPase (TPase-ORF) were extracted by ORFfinder (https://www.ncbi.nlm.nih.gov/orffinder/) and used as input into `bedtools coverage` (Quinlan and Hall 2009) to analyze the breadth of coverage of TPase-ORF. Those showing >90% coverage of the ORF were considered autonomous expression candidates of TIR-TEs.

To compare the expression level of genes co-localizing with autonomous expression candidates with that of housekeeping genes, grapevine *ACTIN* genes were collected according to the Gene Ontology annotation file in Díaz-Riquelme et al. (2016). The FPKM values shown in Figure 3.26 A-C were the average from the replicates of each time point.

Computational scripts used in this chapter can be found in Appendix D.2.

## 3.4 Results

### 3.4.1 The integrity of expression candidates

Based on the reference genome, only 4.4% of all annotated TEs were considered full-length by our criteria outlined above, with all remaining loci being regarded as fragmented and likely non-autonomous elements (Figure 3.4 A). For the untreated embryogenic callus (denoted as T=0; see section 2.3.1), 338 (9.1%) of the 3,698 expression candidates were full length (Figure 3.4 B), which accounted for only 0.16% of the total annotated elements (Table 3.1). The mock treatment, resembling a temporal wounding treatment at the onset of the time-series experiments, resulted in 5,524 expression candidates, of which 9% (497 TE loci) were full-length (Figure 3.4 C, Table 3.1). Continuous incubation with either live yeast (*H. uvarum*) or *Botrytis* (*B. cinerea*) extract revealed, respectively, 5,531 and 5,571 expression candidates, of which 9.7% (539 TE loci) and 9.3% (481 TE loci) were full-length (Figure 3.4 DE, Table 3.1). In the mock, yeast and *Botrytis* treatments, the number of expression candidates was all 50% more than that in T=0. The comparison of the TE loci identity of the four sets expression candidates shows that 2,351 expression candidates were conserved in all four sets, while 4,482 expression candidates were presented in at least one of the three treatments but not in T=0 (Appendix C.2), meaning that mock, yeast and *Botrytis* treatments stimulated transcriptional expression of additional sets of TE loci that were not seen in T=0.

**Table 3.1 All annotated TEs categorized by integrity and transcriptional activity across treatments**

| | T=0 | | Mock | | Yeast treatment | | *Botrytis* treatment | |
|---|---|---|---|---|---|---|---|---|
| | # TEs | Percentage | # TEs | Percentage | # TEs | Percentage | # TEs | Percentage |
| **No expression** | 195,181 | 87.36% | 179,568 | 80.38% | 168,340 | 75.35% | 168,829 | 75.57% |
| **Under threshold** | 24,532 | 10.98% | 38,319 | 17.15% | 49,540 | 22.17% | 49,411 | 22.12% |
| **Fragmented expr. candidates** | 3,360 | 1.50% | 5,027 | 2.25% | 4,992 | 2.23% | 4,690 | 2.10% |
| **Full length expr. candidates** | 338 | 0.16% | 497 | 0.22% | 539 | 0.24% | 481 | 0.21% |
| **Sum (all annotated TEs)** | 223,411 | 100% | 223,411 | 100% | 223,411 | 100% | 223,411 | 100% |

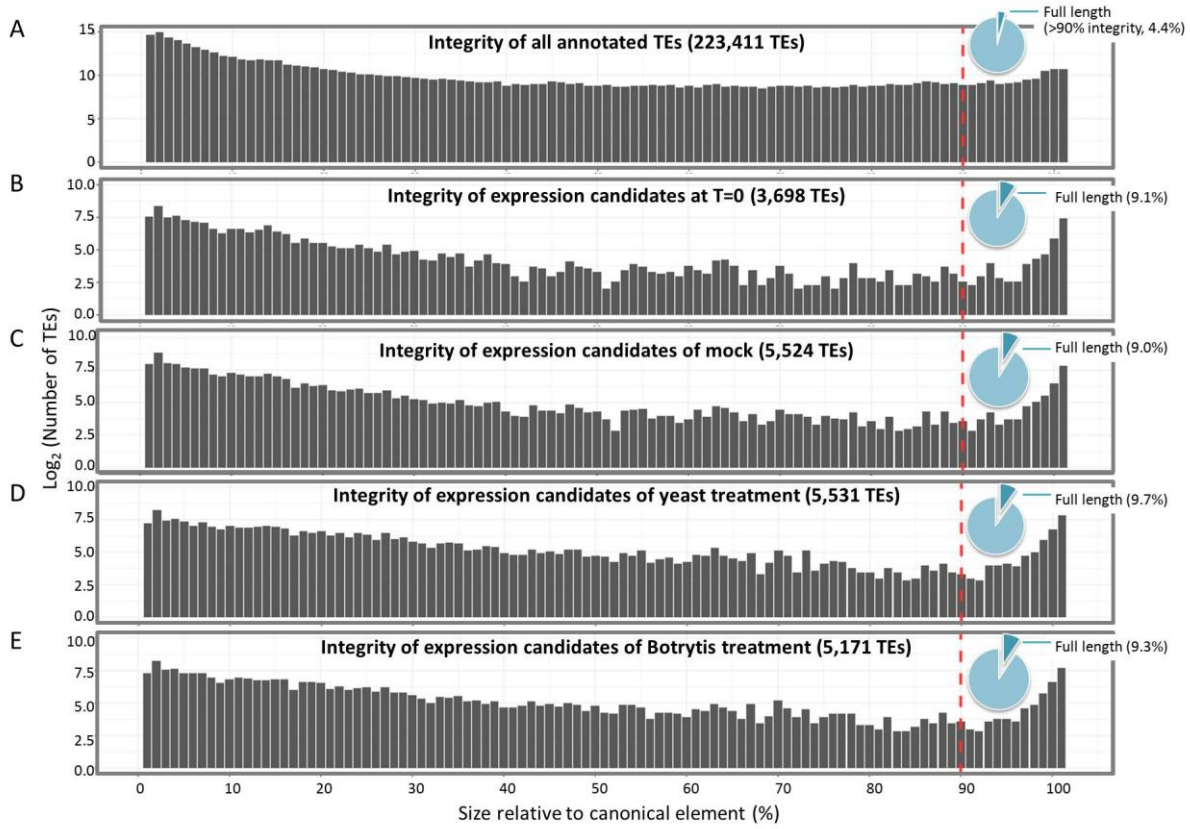



**Figure 3.4 Integrity of annotated TEs**

The length of each individual TE locus was compared with the size of the corresponding canonical element. TEs with over 90% integrity were denoted as full length. **(A)** All annotated TEs (223,411 TEs). **(B-E)** Expression candidates of T=0 (B), mock (C), yeast (D), and *Botrytis* (E) treatments.

### 3.4.2 Survey for the most recently active TE

A further breakdown of the expression candidates binned by TE family and whether candidates are distinguishable with unique-mapping reads can reveal transcriptionally active TE families and those that may have been recently mobilised. The whole collection of expression candidates were pooled from the three sub-pipelines, one of which applies both unique- and multi-mapping reads (sub-pipeline 2), and the other two only adopt unique-mapping reads to quantify expression level (sub-pipeline 1 and 3). Therefore some of the candidates can be tracked by unique-mapping reads while others not (Figure 3.5 A, B); the former is termed 'trackable' expression candidates, and the latter is termed 'untrackable expression candidates (as previously shown in section 2.4.2). As mentioned in chapter 2, the trackable loci are likely to be older insertions having accumulated mutations, whereas the untrackable loci, especially full-length untrackable loci, could be newly transposed elements that are highly similar or identical, hence contributing to the alignment of reads to multiple loci. While TE families showing the highest abundance of expression candidates or sequencing reads might be the most "transcriptionally active" TE family, this family is not necessarily the most "autonomously

mobile" family if its expression candidates are largely fragmented and trackable loci. Hence this section investigates the activity of TE families by deciphering the properties of expression candidates hierarchically as follows:

(1) the abundances of expression candidates of each TE family;

(2) the divergence of expression candidates indicated by the abundances of trackable/un-trackable expression candidates;

(3) the integrity of expression candidates.

For the embryogenic callus at T=0, 2,565 (69%) of the total expression candidates (3,698 TEs) showed evidence of transcription that can be represented by unique-mapping reads (trackable candidates), whereas the remaining 1,280 TEs (untrackable) remained indistinguishable (Figure 3.5 B). These two groups of candidates were further categorized hierarchically by family (Figure 3.5 C) and integrity (Figure 3.5 D). Among the total 232 TE families (corresponding to 9 superfamilies) presented on the y-axis of Figure 3.5 C, 174 of them contained expression candidates at T=0 and thus considered as expressed families. Of these 174 expressed TE families, there are 102 families each contained fewer than 10 expression candidates, and another 32 families each containing 10 to 23 (the median abundances of expression candidates among expressed TE families) expression candidates (Appendix C.3), meaning that most of the expressed TE families didn't have many active TE loci. It is noticeable that 6 TE families had more than 100 expression candidate loci; these families are Copia-23 (211 expression candidates), Gypsy-12 (175 expression candidates), VLINE1 (245 expression candidates), VLINE4 (211 expression candidates), VLINE5 (117 expression candidates) and VLINE6 (139 expression candidates).

With the exception of Copia-23 and Copia-3, the majority of the 174 expressed TE families in T=0 demonstrated fragmented expression candidates that were trackable and lacked un-trackable expression candidates (Figure 3.5 D). Half of the expressed TE families (87 of the 174 expressed families) had zero un-trackable expression candidates, and the other 74 TE families had only 1 to 20 un-trackable expression candidates, most of which are fragmented. This observation demonstrates that the vast majority of the expressed TE families comprised transcriptionally active loci that are mostly fragmented and could be identified by polymorphisms (e.g. SNVs and INDELs). Copia-23 and Copia-3 families were over-represented with expression candidates that were both untrackable and potentially autonomous (full length). These findings were concordant with the observations in mock (Figure 3.6), yeast (Figure 3.7) and *Botrytis* (Figure 3.8) treatments. Although there were more TE families obtaining low numbers of full-length candidates in each of these three treatments than that

at T=0, most of the expressed families are still in short of untrackable expression candidates, and Copia-3 and Copia-23 are still the two families comprising most full-length untrackable expression candidates (Appendix C.3).

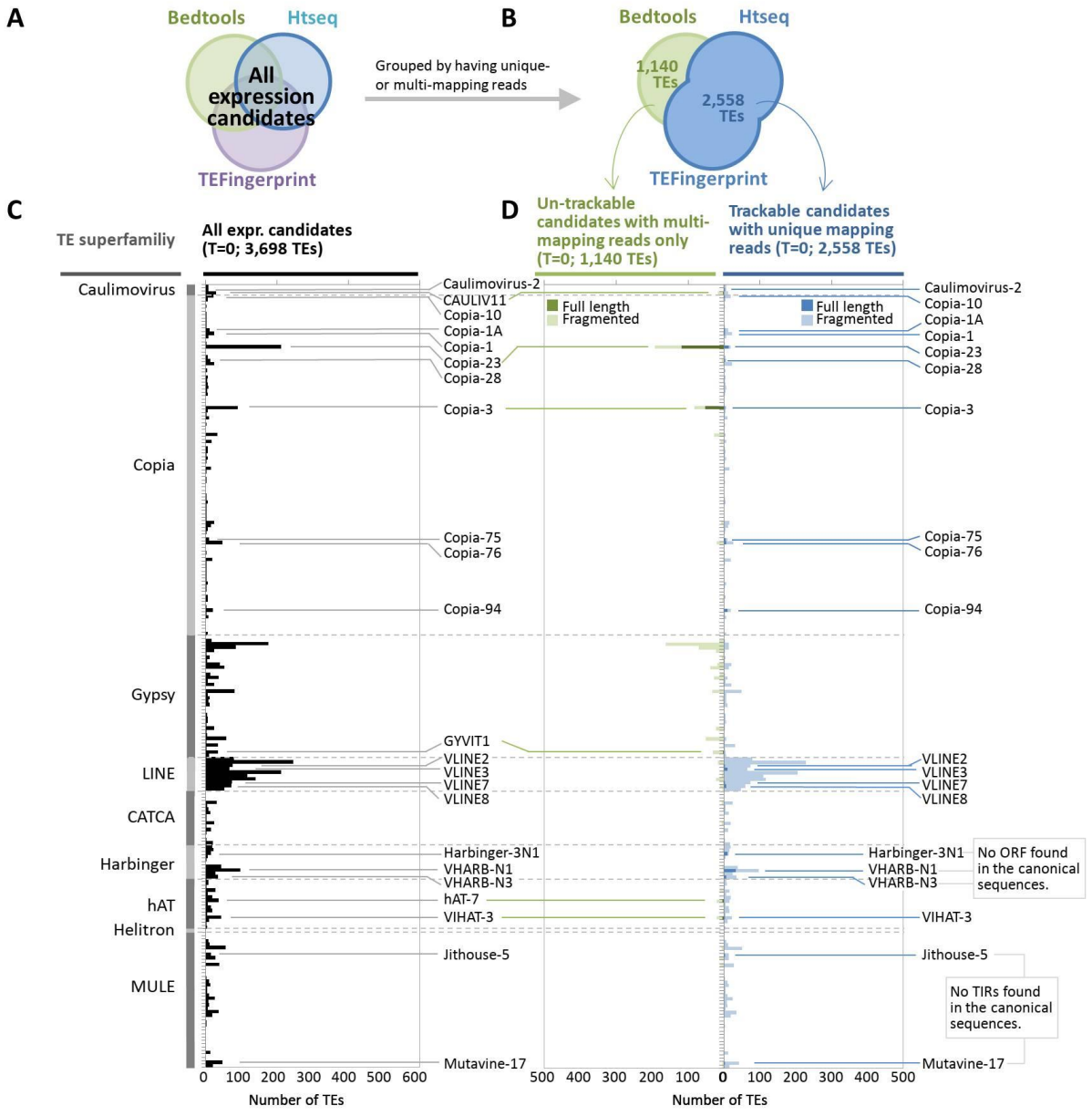



**Figure 3.5 Transcriptionally active TE families at T=0**

**(A, B)** The expression candidates found by the three sub-pipelines (A) can be grouped by the presence (trackable) or absence (un-trackable) of unique-mapping reads (B). **(C, D)** All expression candidates were categorized by families. Each bar represents a TE family containing expression candidates (C). The expression candidates were then further grouped into un-trackable (green) and trackable (blue) candidates (D), those of which full-length were filled with either dark green or dark blue. TE families containing at least two full-length expression candidates of either group were indicated. Note that Harbinger families missing open reading frame (ORF) encoding transposase and MULE families lack of terminal inverted repeats (TIRs) in their canonical sequences were indicated.

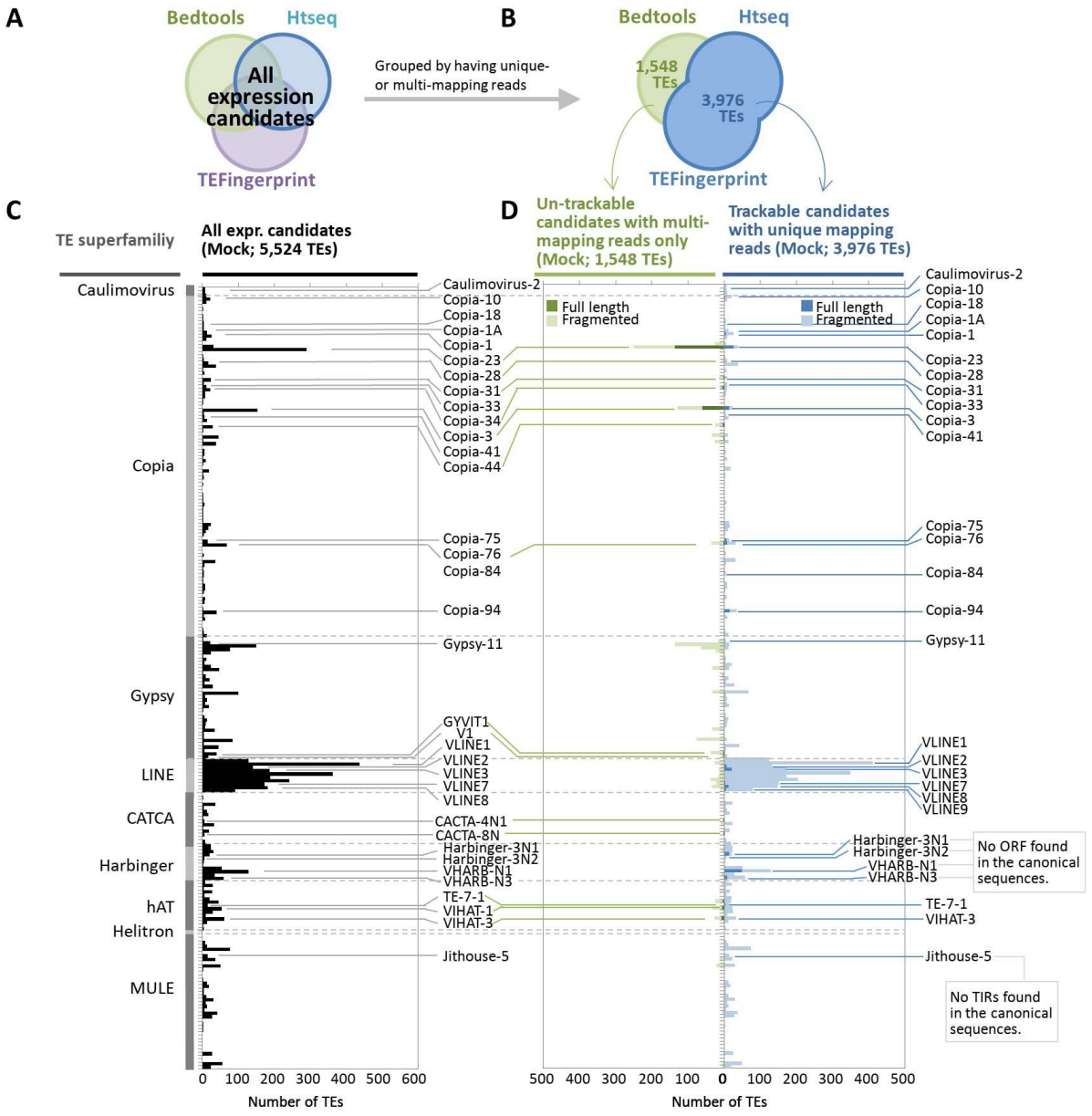



**Figure 3.6 Transcriptionally active TE families in mock treatment**

**(A, B)** In mock treatment, the 5,524 expression candidates found by the three sub-pipelines (A) consisted of 3,976 trackable and 1,548 untrackable expression candidates (B). **(C, D)** All expression candidates were grouped by families (C), followed by a further categorization by the presence (trackable, blue) or absence (untrackable, green) of unique-mapping reads (D). Full-length candidates were filled with either dark green or dark blue. TE families containing at least two full-length expression candidates of either group were indicated. Harbinger families missing ORF encoding transposase and MULE families lack of TIRs in their canonical sequences were indicated.

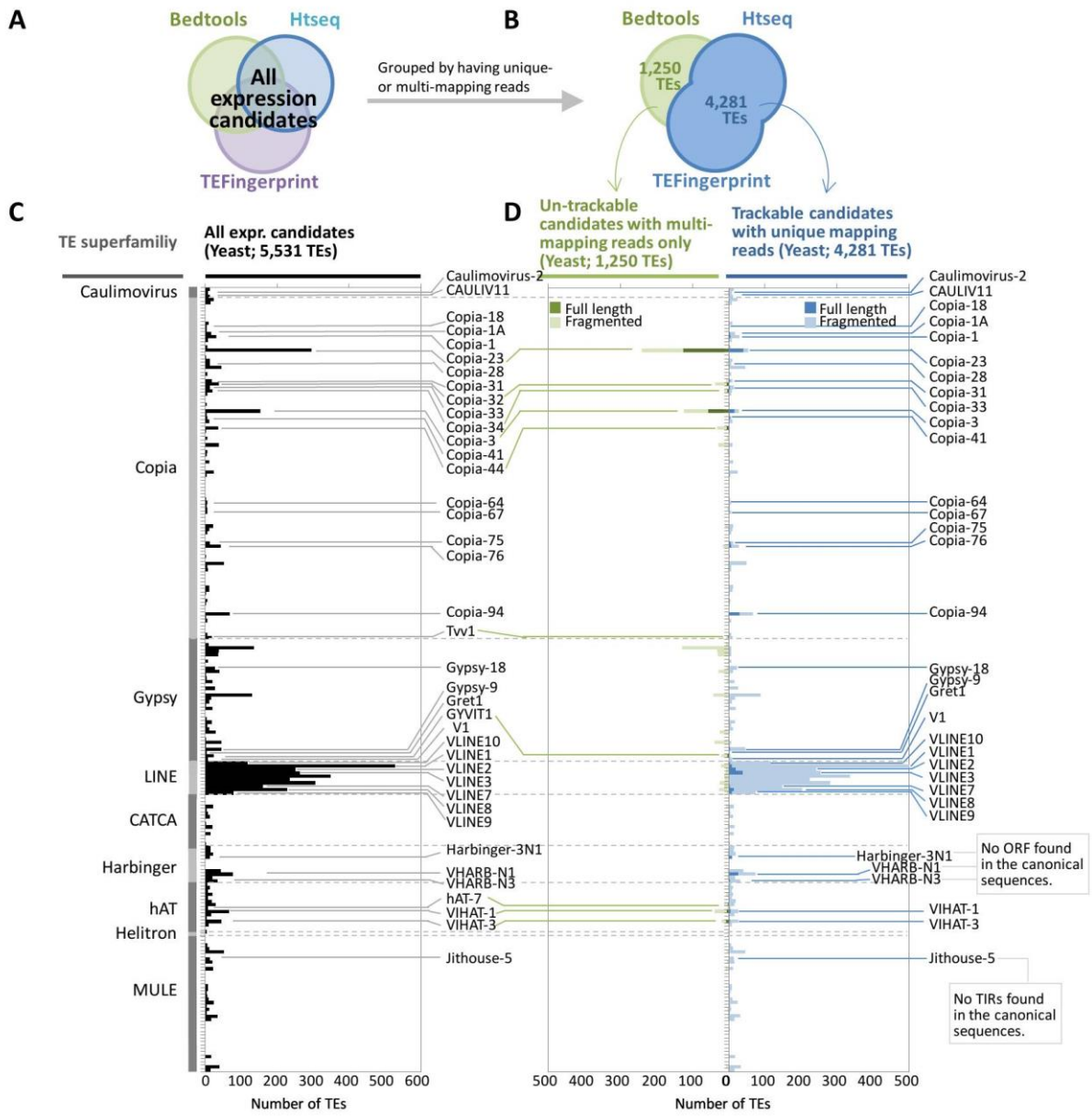



**Figure 3.7 Transcriptionally active TE families in yeast treatment**

**(A, B)** In yeast treatment, the 5,531 expression candidates found by the three sub-pipelines (A) included 4,281 trackable and 1,250 untrackable expression candidates (B). **(C, D)** All expression candidates were categorized by families (C) and then further grouped into either trackable (blue) or untrackable (green) group (D), those of which full-length were denoted by dark blue or dark green accordingly. TE families containing at least two full-length expression candidates of either group were indicated. Harbinger families missing ORF encoding transposase and MULE families lack of TIRs in their canonical sequences were indicated.

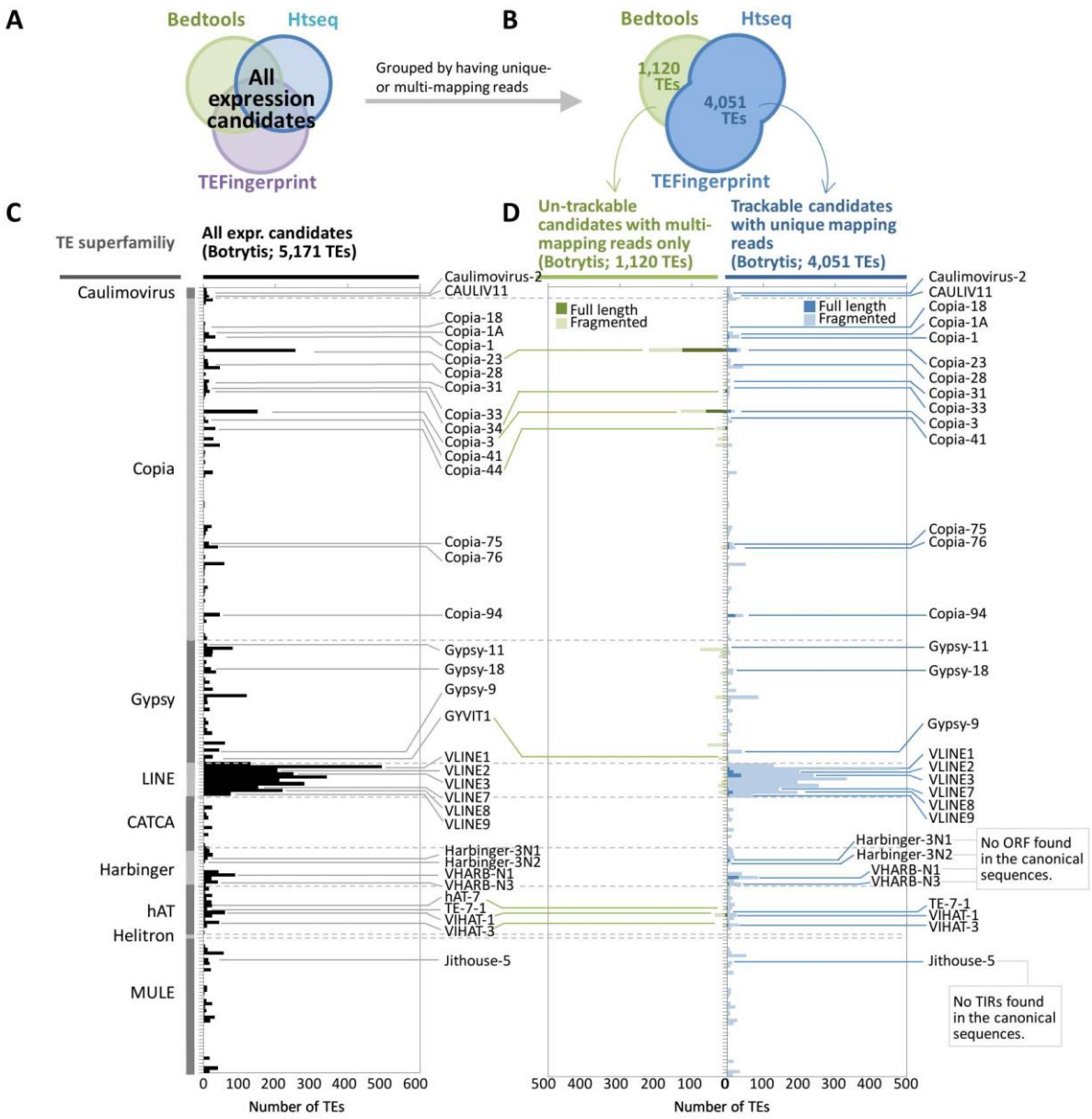



**Figure 3.8 Transcriptionally active TE families in *Botrytis* treatment**

**(A, B)** In *Botrytis* treatment, the 5,171 expression candidates found by the three sub-pipelines (A) were composed of 4,051 trackable and 1,120 untrackable expression candidates (B). **(C, D)** All expression candidates were categorized by families (C) and then further grouped into either trackable (blue) or untrackable (green) group (D), those of which full-length were denoted by dark blue or dark green accordingly. TE families containing at least two full-length expression candidates of either group were indicated. Harbinger families missing ORF encoding transposase and MULE families lack of TIRs in their canonical sequences were indicated.

A closer look at the sequences of the canonical and 90 annotated full-length Copia-3 elements shows a condensed phylogenetic cluster mostly comprised of intact Copia-3 indicated by the presence of LTRs flanking INT domain (Figure 3.9). This cluster included 26 sequences, 19 of which were intact (i.e. structurally autonomous; see section 3.3.3 for explanation) un-trackable candidates with over 90% read coverage of the annotated INT domain, and 4 of which were intact trackable Copia-3 with nearly full transcription of INT. The majority of the remaining un-trackable expression candidates formed three groups, the largest two of which were close to the cluster formed by intact Copia-3 loci. The opposite distal end of the tree was mostly occupied by un-expressed full-length Copia-3 loci.

The neighbour-joining tree built from the canonical and 220 full-length Copia-23 sequences revealed that the 11 untrackable and four trackable intact candidates with nearly full transcription across INT were scattered in 5 broom-like clusters densely packed with other un-trackable candidates that had either lost intact LTRs or lacked full INT coverage (Figure 3.10). These compact clades with short branches were distinguished from the sequences of unexpressed full-length Copia-23.

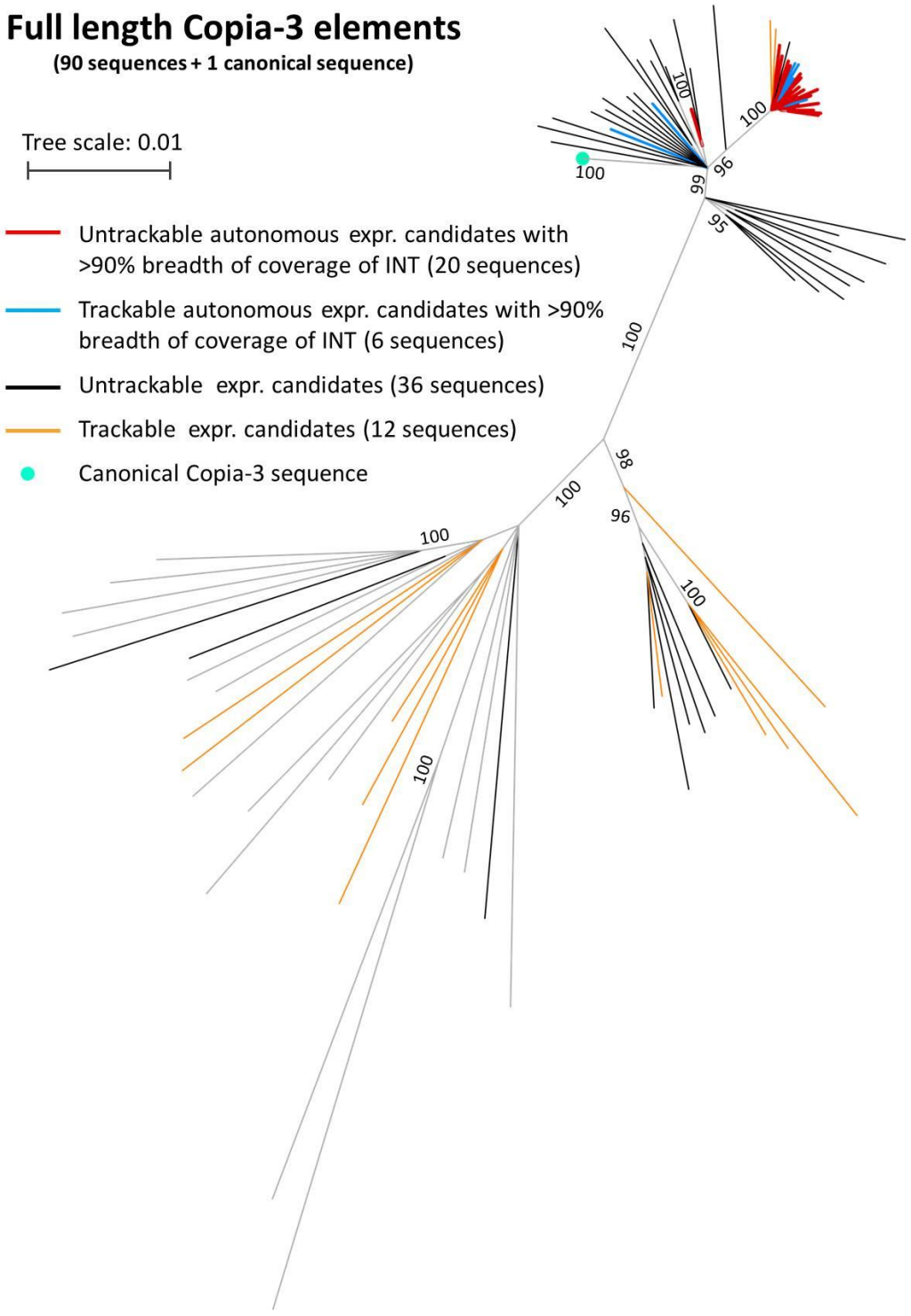



**Figure 3.9 The consensus tree of full-length Copia-3 elements**

Copia-3 elements retaining LTR pairs with >90% INT covered by sequencing reads were considered structurally autonomous expression candidates, those of which have unique-mapping reads (trackable) were coloured blue, and the remaining untrackable ones were denoted by red lines. For the rest of the full-length Copia-3 expression candidates, those can be distinguished by unique-mapping reads were yellow; otherwise, they were coloured black. Grey lines denote branches or non-expression candidates.

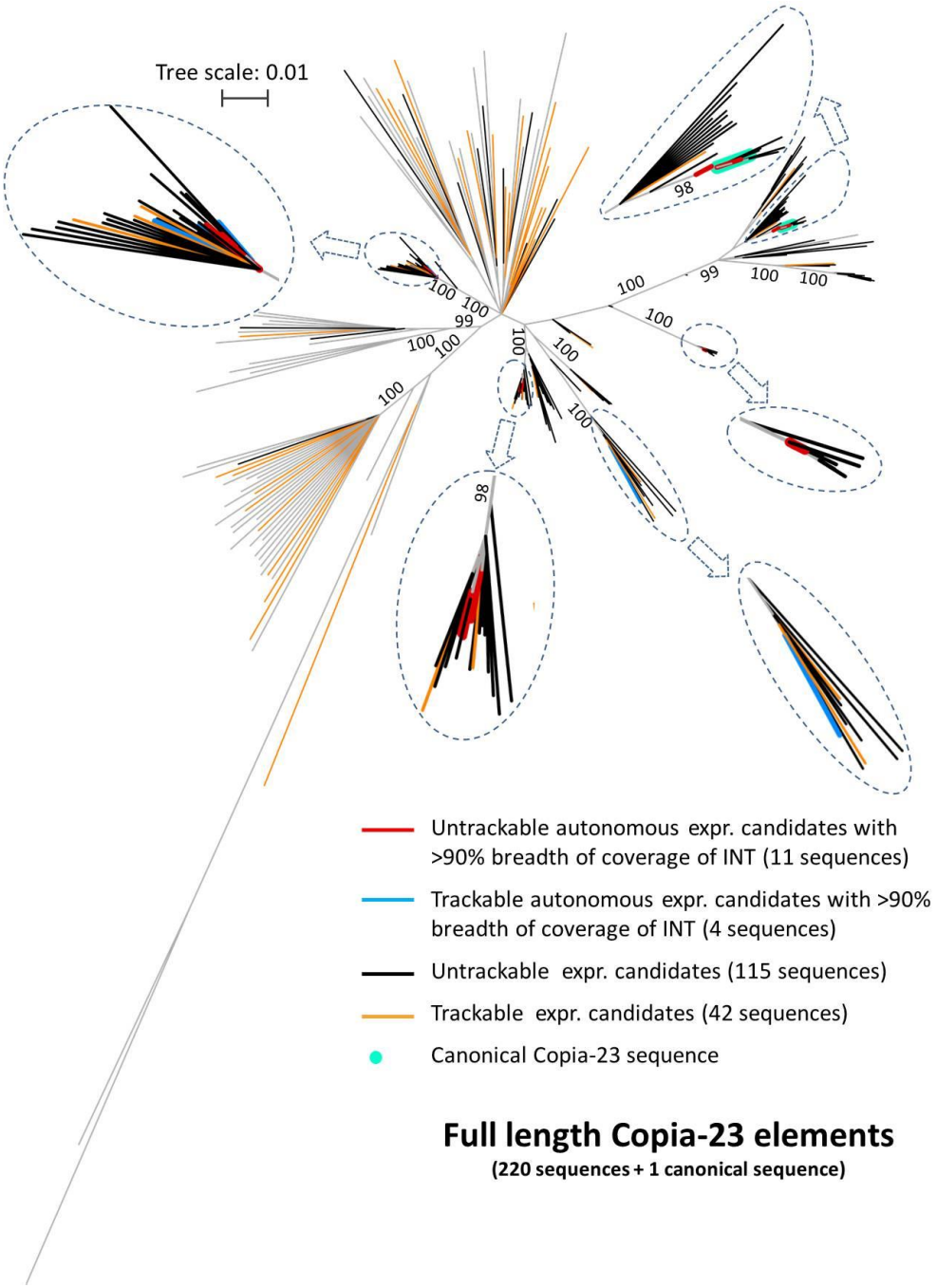



**Figure 3.10 The consensus tree of full-length Copia-23 elements**

Copia-23 elements retaining LTR pairs with >90% INT covered by sequencing reads were considered structurally autonomous expression candidates, those of which trackable were coloured blue and the remaining un-trackable ones were denoted by red lines. For the rest of the full-length Copia-23 expression candidates, those can be distinguished by unique-mapping reads were yellow; otherwise, they were coloured black. Grey lines denote branches or non-expression candidates. Clades containing trackable or un-trackable autonomous expression candidates were indicated with enlarged illustrations.

To test the contribution of the full-length Copia-3 loci to the pool of Copia-3-related transcripts, reads mapping to all Copia-3 candidates were categorized into four groups by whether they mapped to full-length/fragmented and trackable/un-trackable candidates. This analysis revealed that each category contained reads shared with one or more categories, irrespective of treatments and time-point of treatment (Appendix C.4 Figure C.3). Nonetheless, each category obtained a unique subset of reads that only mapped to one of the four groups of expression candidates, meaning none of the group was able to represent the whole collection of Copia-3 transcripts. The same analysis was applied to reads mapping to all Copia-23 expression candidates. This analysis also demonstrated reads shared across different categories and those unique to a single category (Appendix C.4 Figure C.4).

Copia-3 and Copia-23 belong to LTR-retrotransposon (LTR-TEs). This type of TEs is known for the identical long terminal repeat (LTR) at both ends upon insertion. The pair of LTRs gradually accumulate independent mutations across time. The more diverse the two LTR sequences, the more time that has passed since insertion. Therefore to test whether Copia-3 and Copia-23 were active more recently than other LTR-TEs, the divergence of each pair of LTR was analysed. From this analysis, the insertion time of each individual TE was estimated., The insertion dates of the 87 and 177 complete copies (i.e. structurally autonomous loci; see section 3.3.3 for explanation) of Copia-3 and Copia-23, respectively, were calculated based on the divergence of individual pairs of LTRs for each element. The peak of Copia-3 and Copia-23 mobilization was then estimated from the distribution of insertion times and found to occur approximately 0.02 and 0.017 million years ago (MYA), respectively (Figure 3.11 A). Peak insertion times of the other 39 LTR-TE families with at least 10 intact copies were analysed in the same way (Figure 3.11 B, Table 3.2). Most LTR-TE families experienced bursts no longer than 4.5 million years ago (MYA). Note that Copia-3 and Copia-23 were the most recently active LTR-TE families, with their bursts occurred around 0.02 and 0.017 MYA, respectively (Table 3.2). Comparison of the peak insertion time of Copia-3, Copia-23 and the other 5 Copia families, which obtained trackable full-length candidates across all treatments but lacked un-trackable full-length candidates, showed that Copia-3 and Copia-23 experienced significantly more recent bursts than these five other families (Figure 3.11 C).

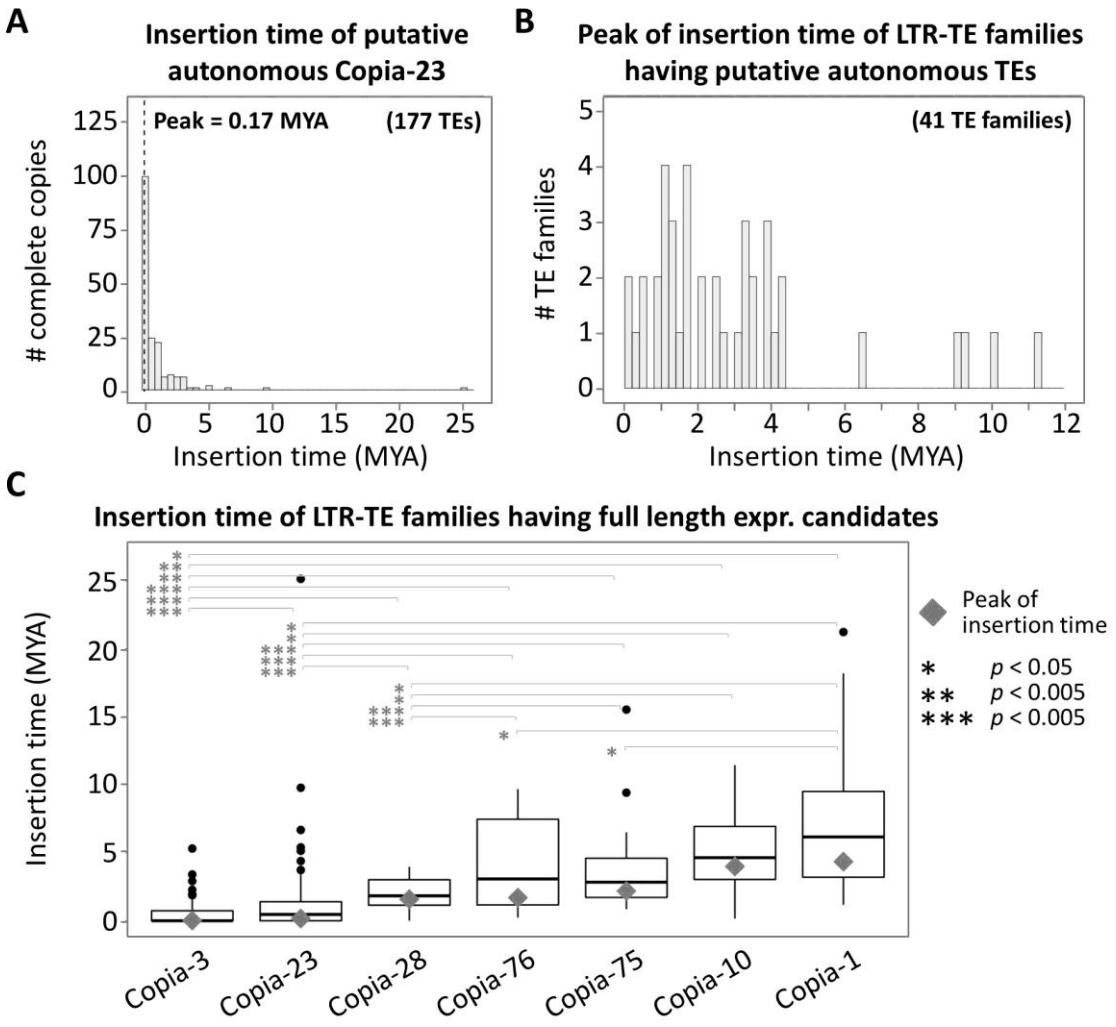



**Figure 3.11 Insertion dates of LTR-TE families containing complete copies**

**(A)** Distribution of the insertion dates for 177 complete copies of the Copia-23 family. The peak of amplification (0.17 MYA) is indicated as the dashed line. **(B)** Distribution of the insertion dates using 41 LTR-TE families with at least 10 complete copies. **(C)** Comparison of the transposition burst time among LTR-TE families containing full-length expression candidates. The families were ordered by the peak of insertion time (grey diamonds). The asterisks denote a significant level of t-test for the mean of insertion time as indicated.

**Table 3.2 Summary of intact LTR-TE families with intact individual TE loci (complete copies) and the peak of insertion time**

| TE family | # copy | # full-length copy | # complete copy | Peak of insertion (MYA) |
|---|---|---|---|---|
| Copia-3 | 403 | 90 | 87 | 0.019322 |
| Copia-23 | 645 | 220 | 177 | 0.166211 |
| Copia-12 | 183 | 14 | 13 | 0.243755 |
| Copia-70 | 183 | 17 | 16 | 0.454001 |
| Gypsy-16 | 501 | 15 | 15 | 0.49476 |
| Copia-24 | 109 | 10 | 10 | 0.964731 |
| Copia-44 | 249 | 12 | 12 | 0.993348 |
| Copia-38 | 91 | 16 | 14 | 1.038244 |
| Copia-50 | 145 | 14 | 13 | 1.168439 |
| Gypsy-V1 | 413 | 49 | 49 | 1.181391 |
| Gypsy-GYVIT1 | 1026 | 60 | 59 | 1.199088 |
| Gypsy-Gret1 | 467 | 65 | 54 | 1.205757 |
| Gypsy-11 | 1148 | 17 | 16 | 1.274305 |
| Gypsy-9 | 1160 | 55 | 54 | 1.28101 |
| Copia-28 | 141 | 18 | 16 | 1.584873 |
| Copia-76 | 367 | 29 | 18 | 1.69977 |
| Copia-41 | 138 | 10 | 10 | 1.716931 |
| Gypsy-14 | 1219 | 12 | 10 | 1.779129 |
| Gypsy-6 | 675 | 23 | 22 | 1.798286 |
| Copia-34 | 342 | 28 | 24 | 2.120972 |
| Copia-75 | 296 | 34 | 31 | 2.186037 |
| Gypsy-3 | 3052 | 79 | 76 | 2.520534 |
| Copia-40 | 348 | 18 | 18 | 2.524404 |
| Copia-7 | 234 | 17 | 12 | 2.759228 |
| Copia-94 | 473 | 117 | 32 | 3.127911 |
| Gypsy-4 | 1311 | 20 | 19 | 3.204009 |
| Gypsy-20 | 1306 | 15 | 13 | 3.218455 |
| Copia-33 | 1162 | 94 | 87 | 3.260797 |
| Copia-67 | 149 | 22 | 14 | 3.491503 |
| Gypsy-32 | 74 | 13 | 13 | 3.527549 |
| Copia-99 | 44 | 20 | 10 | 3.868751 |
| Copia-18 | 188 | 22 | 17 | 3.880459 |
| Copia-10 | 7156 | 27 | 27 | 3.995773 |
| Copia-31 | 921 | 127 | 120 | 4.146948 |
| Copia-1 | 696 | 20 | 19 | 4.346034 |
| Gypsy-33 | 676 | 18 | 18 | 4.387968 |
| Copia-64 | 205 | 12 | 11 | 6.480961 |
| Copia-86 | 330 | 16 | 14 | 9.118008 |
| Gypsy-19 | 3404 | 11 | 11 | 9.20991 |
| Copia-53 | 923 | 19 | 19 | 10.15459 |
| Copia-83 | 518 | 18 | 18 | 11.33491 |

### 3.4.3 Hierarchical classifications of expression candidates by location, integrity, and distinctness

In order to investigate whether there is location bias for all annotated TEs and expression candidates, the annotated genome was compartmented into genic and intergenic regions. The former comprised gene units, which were made of exons and introns included from the transcription start sites to the transcription stop sites of genes, and flanking regions of genes which encompassed 2kb upstream (N-flanks) and 2kb downstream (C-flanks) of corresponding translation start and stop sites (Figure 3.12 A).

All annotated TEs intersected with specific genome compartments were categorized accordingly and hierarchically in the order of genic/intergenic regions, location within the genic region (e.g. exon, intron, flanks), and integrity (full-length or fragmented). Over half of all annotated TE loci fell into intergenic regions (126,976 TEs, 56.83%), while 96,435 (43.16%) TEs co-localized with genes (Figure 3.12 B). About half of the genic TEs were in flanking regions with no particular preference for either flank. As expected, intronic TEs comprised the majority of TEs in gene units (Figure 3.12 B, Table 3.3).

Expression candidates were classified in the same way, with additional categories added, including the transcriptional activity of co-localized genes (i.e. TEs associated with expressed or non-expressed genes) and the presence or absence of unique-mapping reads (trackable or un-trackable). Untreated embryogenic callus (T=0), as well as mock, yeast and Botrytis treatments, respectively, showed 71.47%, 75.69%, 74.62%, and 76.77% of the expression candidates located in the genic regions (Figure 3.12 C-F, Table 3.4). Delving deeper into the insertion context, about two-thirds of the genic TE expression candidates overlapped with introns; in particular, there was a bias toward insertion into introns of expressed genes (Figure 3.12 C-F, Table 3.4).

In all location categories (exon, intron, flanks, and intergenic regions), fragmented expression candidates were more prevalent than full-length expression candidates in T=0, mock, yeast and *Botrytis* treatments (Figure 3.12 C-F). Trackable expression candidates had contributed to a substantial proportion of expression candidates in each category shown in Figure 3.12 C-F. The overall ratio of trackable versus un-trackable expression candidates is likely to differ depending on treatments, and thus an X-square test (shown in the next paragraph) was conducted on this matter to statistically examine the differences.

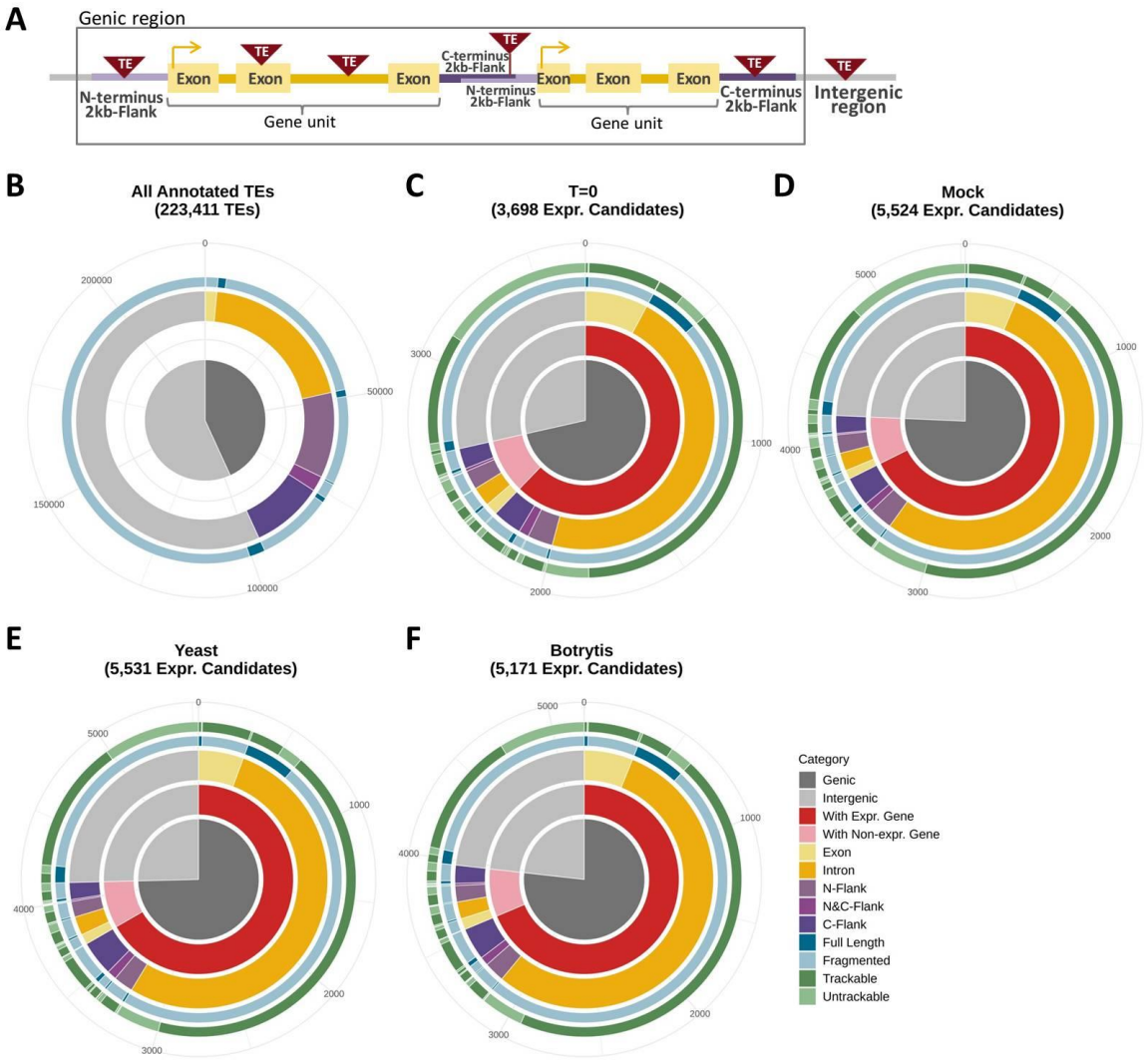



**Figure 3.12 Hierarchical classifications of expression candidates by location, integrity, and distinctness.**

**(A)** TEs overlapping with exon, intron, or 2kb upstream (N-terminus) or downstream (C-terminus) of a gene were denoted as genic TEs; otherwise, they were grouped as intergenic TEs. **(B)** All annotated TEs were categorized hierarchically by region (centre), location (internal layer) and integrity (outer-most layer). **(C-F)** Expression candidates of each treatment were categorized in the order of region (centre), the transcriptional activity of co-localized genes (2$^{nd}$ layer), location (3$^{rd}$ layer), integrity (4$^{th}$ layer), and the presence/absence of unique-mapping reads (outer-most layer).

**Table 3.3 Hierarchical categorization of *V. vinifera* annotated TEs by location and integrity.**

Numbers of TEs (#TE) in black sum up to 223,411 annotated TEs, while the corresponding percentages (Perc.) in black sum up to 100.00%. Subtotals of indicated categories were denoted in grey.

| Region | Location | | All annotated TEs | |
|---|---|---|---|---|
| | | | #TE | Perc. |
| **Genic** | Gene unit | Exon | 3,257 | 1.46% |
| | | Intron | 44,827 | 20.06% |
| | *Subtotal (Gene unit)* | | *48,084* | *21.52%* |
| | Flanks | N-Flank | 23,895 | 10.70% |
| | | N&C-Flank | 4,258 | 1.91% |
| | | C-Flank | 20,198 | 9.04% |
| | *Subtotal (Flanks)* | | *48,351* | *21.64%* |
| *Subtotal (Genic)* | | | *96,435* | *43.16%* |
| **Intergenic** | | | 126,976 | 56.84% |
| *Subtotal (Intergenic)* | | | *126,976* | *56.84%* |
| **Sum** | | | 223,411 | 100.00% |

**Table 3.4 Hierarchical categorization of expression candidates by location and integrity.**

Numbers of TEs (#TE) in black sum up to 223,411 annotated TEs, while the corresponding percentages (Perc.) in black sum up to 100.00%. Subtotals of indicated categories were denoted in grey.

| Region | Gene activity | Location | | T=0 | | Mock | | Yeast | | Botrytis | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | #TE | Perc. | #TE | Perc. | #TE | Perc. | #TE | Perc. |
| **Genic** | With expr. Gene | Gene unit | Exon | 290 | 7.84% | 338 | 6.12% | 312 | 5.64% | 301 | 5.82% |
| | | | Intron | 1,712 | 46.30% | 2,654 | 48.04% | 2,933 | 53.03% | 2,626 | 50.78% |
| | | *Subtotal (Gene unit)* | | *2,002* | *54.14%* | *2,992* | *54.16%* | *3,245* | *58.67%* | *2,927* | *56.60%* |
| | | Flanks | N-Flank | 117 | 3.16% | 449 | 8.13% | 136 | 2.46% | 333 | 6.44% |
| | | | N&C-Flank | 46 | 1.24% | 83 | 1.50% | 65 | 1.18% | 76 | 1.47% |
| | | | C-Flank | 135 | 3.65% | 163 | 2.95% | 242 | 4.38% | 191 | 3.69% |
| | | *Subtotal (Flanks)* | | *298* | *8.06%* | *695* | *12.58%* | *443* | *8.01%* | *600* | *11.60%* |
| | *Subtotal (With expr. gene)* | | | *2,300* | *62.20%* | *3,687* | *66.75%* | *3,688* | *66.68%* | *3,527* | *68.21%* |
| | With non-expr. Gene | Gene unit | Exon | 56 | 1.51% | 88 | 1.59% | 71 | 1.28% | 96 | 1.86% |
| | | | Intron | 90 | 2.43% | 88 | 1.59% | 129 | 2.33% | 110 | 2.13% |
| | | *Subtotal (Gene unit)* | | *146* | *3.95%* | *176* | *3.19%* | *200* | *3.62%* | *206* | *3.98%* |
| | | Flanks | N-Flank | 89 | 2.41% | 126 | 2.28% | 109 | 1.97% | 106 | 2.05% |
| | | | N&C-Flank | 13 | 0.35% | 70 | 1.27% | 12 | 0.22% | 14 | 0.27% |
| | | | C-Flank | 95 | 2.57% | 122 | 2.21% | 118 | 2.13% | 117 | 2.26% |
| | | *Subtotal (Flanks)* | | *197* | *5.33%* | *318* | *5.76%* | *239* | *4.32%* | *237* | *4.58%* |
| | *Subtotal (With non-expr. gene)* | | | *343* | *9.28%* | *494* | *8.94%* | *439* | *7.94%* | *443* | *8.57%* |
| *Subtotal (Genic)* | | | | *2,643* | *71.47%* | *4,181* | *75.69%* | *4,127* | *74.62%* | *3,970* | *76.77%* |
| **Intergenic** | Intergenic | | | 1,055 | 28.53% | 1,343 | 24.31% | 1,404 | 25.38% | 1,201 | 23.23% |
| *Subtotal (Intergenic)* | | | | *1,055* | *28.53%* | *1,343* | *24.31%* | *1,404* | *25.38%* | *1,201* | *23.23%* |
| **Sum** | | | | 3,698 | 100.00% | 5,524 | 100.00% | 5,531 | 100.00% | 5,171 | 100.00% |

Using the goodness of fit X-square test, the proportion of genic TEs was significantly elevated from 43% of the 'default' distribution of all annotated TEs to 71% of T=0 expression candidates (see the comparison of 'All annotated' versus T=0 in Figure 3.13 A). This location bias is further enhanced in mock, yeast and *Botrytis* treatments, as the genic proportion of expression candidates was significantly increased from 71% in T=0 to 76%, 75% and 77% in mock, yeast and *Botrytis* treatments, respectively.

Among the genic TE loci, the intragenic (i.e. gene-unit) proportion was significantly increased from 50% of the 'default' distribution of total annotated TEs to 81%, 83%, 84%, and 83%, respectively, in T=0, mock, yeast, and Botrytis treatments (Figure 3.13 B).

For expression candidates co-localized with genes (including expression candidates in gene units and flanking regions), the expected number of expression candidates with expressed genes was established based on the proportion of expressed genes (FPKM > 1) relative to the total annotated genes. The statistical test revealed that, in T=0, the observed proportion (87%) of expression candidates co-localized with expressed genes was significantly higher than the expected proportion (47%; Figure 3.13 C). Likewise, in mock, yeast and Botrytis treatments, the proportion of expression candidates co-localized with expressed genes significantly deviated from the expected 50% to the observed 89% (Figure 3.13 C). In addition, this observed proportion (89%) in mock, yeast, and *Botrytis* treatments was significantly 2% higher than the observed proportion (87%) in T=0.

While the presence of external stressors (i.e. wound-like mock treatment, live yeast cultures and *Botrytis* cell extracts) significantly enhanced the aforementioned location bias of expression candidates in untreated embryogenic callus (T=0), these stressors also significantly elevated the proportion of trackable expression candidates from 69% in T=0 to 72%, 77% and 78% in mock, yeast, and *Botrytis* treatments, respectively (Figure 3.13 E). Nonetheless, these treatments didn't significantly change the low proportion (9%) of full-length expression candidates (Figure 3.13 D).

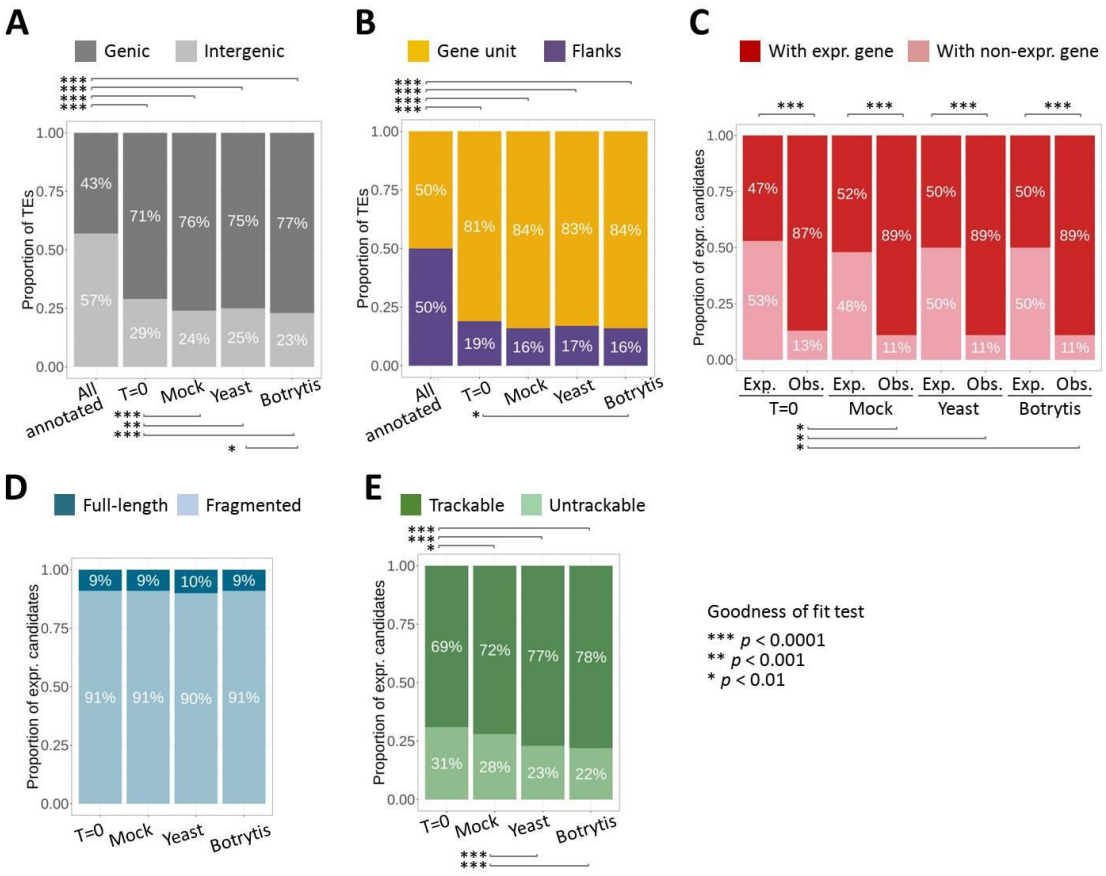



**Figure 3.13 Characteristics of expression candidates in terms of location, integrity and distinctness.**

**(A)** Categorization of annotated TEs and expression candidates by genic/intergenic regions. **(B)** Categorization of annotated genic TEs and genic expression candidates by location relative to genes. **(C)** Classification of genic expression candidates by the transcriptional activity of co-localized genes and statistical comparison between the expected and observed values. **(D-E)** Categorization of all expression candidates by integrity (D) and distinctness (E). The goodness of fit test was performed pair-wisely. All the comparisons reached $p < 0.01$ were labelled. Levels of statistical significance were as indicated. Exp., expected; Obs., observed.

The proposed preference for expressed TE loci to be located in genic regions may be explained by either a general increase in the proportion of genic expression candidates from most TE families or simply a reflection of the genic-enriched annotation of a few TE families that largely contribute to the pool of expression candidates. To test these two assumptions, the genic and intergenic proportions of annotated TE loci and expression candidates were plotted for families belonging to Copia, Gypsy, LINE, hAT and MULE, the five super-families that contributed to the majority of expression candidates (Appendix C.5 Figure C.5-Figure C.9). This analysis first looked at the genic and intergenic proportion of annotated TE loci grouped by families and then categorized expression candidates, in the same way, to examine that whether the genic proportion of expression candidates is higher than that of annotated TE loci in most of the investigated TE families. For the genic and intergenic distribution of annotated TE loci belonging to Copia families, while the genic and intergenic fractions

of annotated TE loci vary by families, about two-thirds of the families show underrepresentation (<50%) of genic TE loci (Appendix C.5 Figure C.5 A). When it comes to expression candidates in T=0, 58 of the 71 expressed families demonstrated higher genic proportions of expression candidates (Appendix C.5 Figure C.5 B) than that of the annotated TE loci (Appendix C.5 Figure C.5 A), meaning that the tendency of expressed TE loci to be located in the genic region is broadly presented in Copia families in T=0. This trend in Copia families is also observed in mock, yeast and *Botrytis* treatments (Appendix C.5 Figure C.5 C-E). The analysis for TE families of Gypsy, LINE, hAT and MULE (Appendix C.5 Figure C.6-Figure C.9) is concordant with the aforementioned findings in Copia; although there are different degrees of elevation in genic proportion, the distribution bias of expression candidates towards genic region seems to broadly happen in most of the families. The same analysis for the location preference within the genic region (exon, intron, and flanking regions) also reveals that the elevation of the intronic fraction of expression candidates is widely presented in most of the TE families (Appendix C.6).

### 3.4.4 Identification of potential origins of autonomous TE transcripts from the short-read RNAseq data

To explore the putative autonomous TE loci that are likely to produce full-length transcripts that are required for autonomous mobilization (hereinafter 'autonomous transcripts'), the sequence integrity and structure of TE loci were firstly screened for competent TE loci annotated in the genome. The competent TE loci that were identified as expression candidates in the stress treatments were then examined for the breadth of read coverage against the TE sequence compartments whose transcription is required for autonomous mobilization (see section 3.2.1.3 and 3.3.8).

The *V. vinifera* reference genome obtains 2,043 full-length LTR-type transposable elements (LTR-TE), of which 1,680 (82.23%) retain LTRs at both ends (Table 3.5). The same filtering approach identified 181, 234, 242, and 220 intact LTR-TE loci in the pools of expression candidates in T=0, mock, yeast, and *Botrytis* treatments, respectively. These expression candidates were further inspected for greater than 90% breadth of coverage across the INT domain, revealing seven expression candidates that are likely to be the origins of autonomous LTR-TE transcripts at T=0, as well as 28, 38 and 28 autonomous expression candidates, respectively, in mock, yeast and *Botrytis* treatments (Table 3.5).

Although there are 23,447 TE loci derived from type I non-LTR retrotransposons (non-LTR-TEs) in the reference genome (Table 2.1), only 179 of these TE loci are full-length non-LTR-TEs (Table 3.6). Among these 179 elements, 159 appear to be potentially autonomous elements based on the criteria outlined above, including those containing full reverse transcriptase (RT) domains in the canonical sequences (Table 3.6). Although in T=0, mock, yeast and *Botrytis* treatments, there were respectively

9, 15, 30, and 30 full-length expression candidates identified as competent loci of non-LTR-TE families, only a LINE expression candidate in mock treatment was nearly fully covered by sequencing reads.

**Table 3.5 Selection of expression candidates potentially producing autonomous Type I LTR-TE transcripts.**

| TE subsets | Treatments | # Selected TEs | | |
|---|---|---|---|---|
| | | Step 1: Full-length | Step 2: Full-length with LTRs | Step 3: >90% INT coverage |
| **Annotated TEs** | | 2,043 → | 1,680 | - |
| **Expr. candidates** | T=0 | 237 → | 181 → | 7 |
| | Mock | 332 → | 234 → | 28 |
| | Yeast | 357 → | 242 → | 38 |
| | Botrytis | 313 → | 220 → | 28 |

**Table 3.6 Selection of expression candidates potentially producing autonomous Type I non-LTR-TE transcripts.**

The competent family denotes those retaining intact reverse transcriptase (RT) domain with putative active sites in the canonical sequence.

| TE subsets | Treatments | # Selected TEs | | |
|---|---|---|---|---|
| | | Full-length | Full-length TE of competent family | >90% coverage |
| **Annotated TEs** | | 179 → | 159 | - |
| **Expr. candidates** | T=0 | 21 → | 9 → | 0 |
| | Mock | 48 → | 15 → | 1 |
| | Yeast | 96 → | 30 → | 0 |
| | Botrytis | 87 → | 30 → | 0 |

**Table 3.7 Selection of expression candidates potentially producing autonomous Type-II TIR-TE TPase transcripts.**

The competent family denotes those retaining terminal inverted repeats (TIRs) and open reading frame (ORF) of intact transposase (TPase) catalytic domain in the canonical sequence.

| TE subsets | Treatments | # Selected TEs | | |
|---|---|---|---|---|
| | | Full-length | Full length TEs of competent family | >90% TPase ORF coverage |
| **Annotated TEs** | | 7,465 → | 29 → | - |
| **Expr. candidates** | T=0 | 74 → | 4 → | 0 |
| | Mock | 114 → | 5 → | 0 |
| | Yeast | 82 → | 3 → | 0 |
| | Botrytis | 76 → | 2 → | 0 |

Almost seven-and-a-half thousand annotated TIR-type DNA transposons (TIR-TEs) were full-length, yet only 29 were associated with the canonical sequences retaining TIRs and an open reading frame (ORF) encoding TPase (Table 3.7). Despite the fact that potentially autonomous expression candidates of TIR-TE families were found in all treatments, none of the TPase ORF of these expression candidates was fully transcribed (Table 3.7).

Focusing on the autonomous LTR-TE expression candidates demonstrating putative full transcription of INT domain (Table 3.5), the seven autonomous expression candidates at T=0 were all found in each of the three other treatments (Figure 3.14 A). There are other 14 autonomous expression candidates shared by mock, yeast and *Botrytis* treatments (Figure 3.14 A). Each pair of the stress treatment shared 1 to 3 autonomous expression candidates (Figure 3.14 A). The yeast treatment showed the most distinct autonomous origins (14 TE loci) that were potentially fully transcribed, while mock and *Botrytis* treatments were uniquely associated with other 2 and 3 autonomous candidates, respectively (Figure 3.14 A). All of the potential loci from which autonomous LTR-TE transcripts may derive (46 TEs) showed a positional preference of insertion in the introns of expressed genes (Figure 3.14 B). Classification by family revealed that Copia-23 and Copia-3 loci were over-represented in this collection (Figure 3.14).

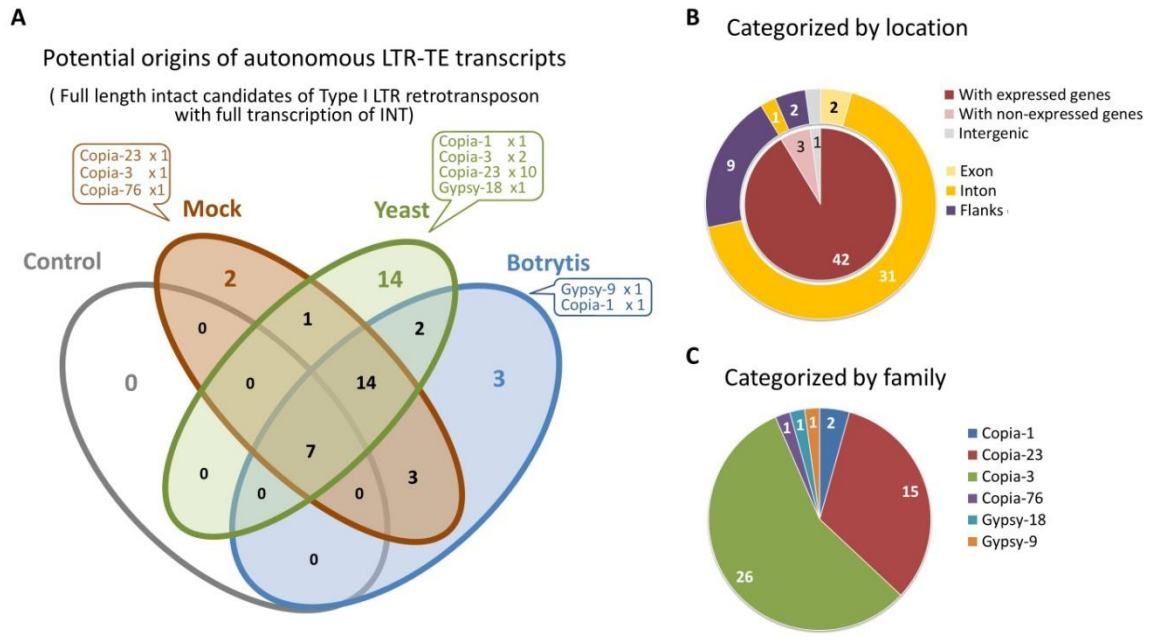



**Figure 3.14 Identification of putative autonomous expression candidates transcriptionally responsive to different stress treatments.**

**(A)** Different sets of potential origins of autonomous LTR-TE transcripts. **(B-C)** Categorization of the potential autonomous expression candidates by location (B) and family (C).

A closer look at the transcriptional activity of these 46 autonomous expression candidates and the co-localized genes revealed only partial consistency between the transcriptional activity of the two (Figure 3.15). Forty-two of these 46 TE loci were identified as expression candidates in all four conditions. Except the 7 TE loci that showed the potential of full-length transcription across all conditions (indicated by rows with all four red blocks in Panel I, Figure 3.15), the majority of these expression candidates were only found to be fully transcribed (i.e. full-length transcription) in some of the treatments (indicated by rows with mixed pink and red blocks in Panel I, Figure 3.15). Nonetheless, nearly all of their co-localized genes were expressed irrespective of treatments (four red blocks across rows in Panel III, Figure 3.15). Only 9 of these genes were differentially expressed in the treatment, where the expression candidates were found potentially fully transcribed (Panel IV, Figure 3.15). Besides, there appeared no particular association of orientation between these TEs and co-localized genes (Panel V, Figure 3.15).

To test whether the genes co-localized with the 46 autonomous TE expression candidates were highly expressed, the FPKM level of these genes was compared with that of the housekeeping genes *ACTIN* and the FPKM quantiles estimated from all expressed genes (Figure 3.16). Although four of the six genes encoding Actin in *V. vinifera* were generally expressed at the level lower than the 3$^{rd}$ quantile in all stress treatments over time, the other 2 *ACTIN* genes were always above the 3$^{rd}$ quantile, indicating the transcriptional level required for producing the housekeeping protein Actin (Figure 3.16 A-C). For the 28 expressed genes associated with autonomous TE expression candidates in mock treatment, only 3 of them showed FPKM values exceeded the 3$^{rd}$ quantile in at least one time point (Figure 3.16 A, D). Likewise, respectively, in yeast and *Botrytis* treatments, there were 38 and 28 expressed genes co-localised with the autonomous TE expression candidates, but only 5 and 4 of these genes were expressed above the 3$^{rd}$ FPKM quantile (Figure 3.16 B-C, E-F). These findings mean that the 46 autonomous TE expression candidates are generally associated with transcriptionally active genes, but these genes are less likely to be highly expressed.

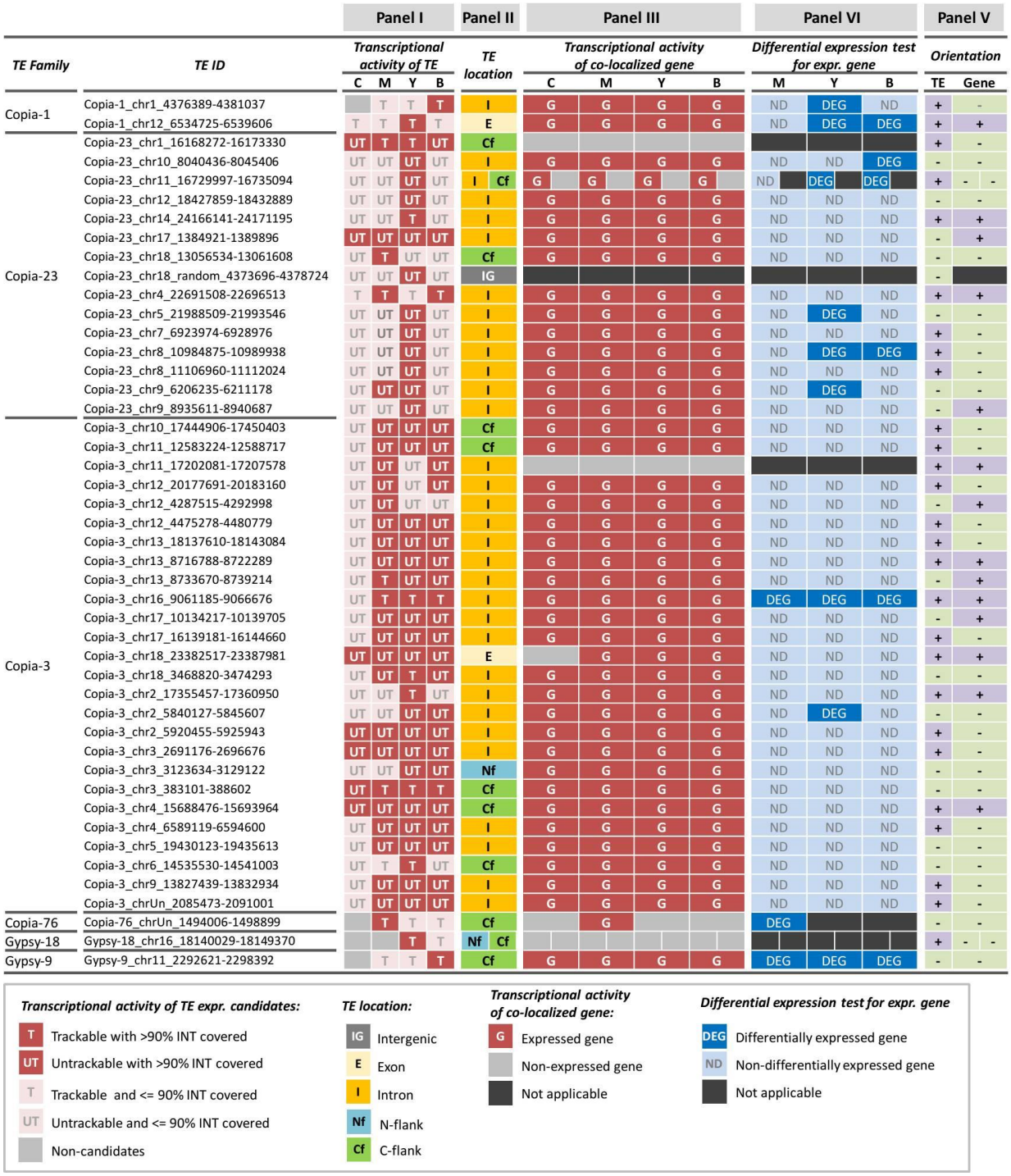



**Figure 3.15 Association of the putative autonomous expression candidates of LTR retrotransposon and the co-localized genes.**

The transcriptional activity of the 46 autonomous expression candidates and co-localized genes were shown in Panel I and III, respectively, with their orientation indicated in Panel V. The location of these TEs was shown in Panel II. The differential expression tests of each expressed genes were as indicated on Panel IV. Meanings of the colour blocks with abbreviations were explained at the bottom. Abbreviation for treatments: C, T=0; M, mock; Y, yeast; B, Botrytis.

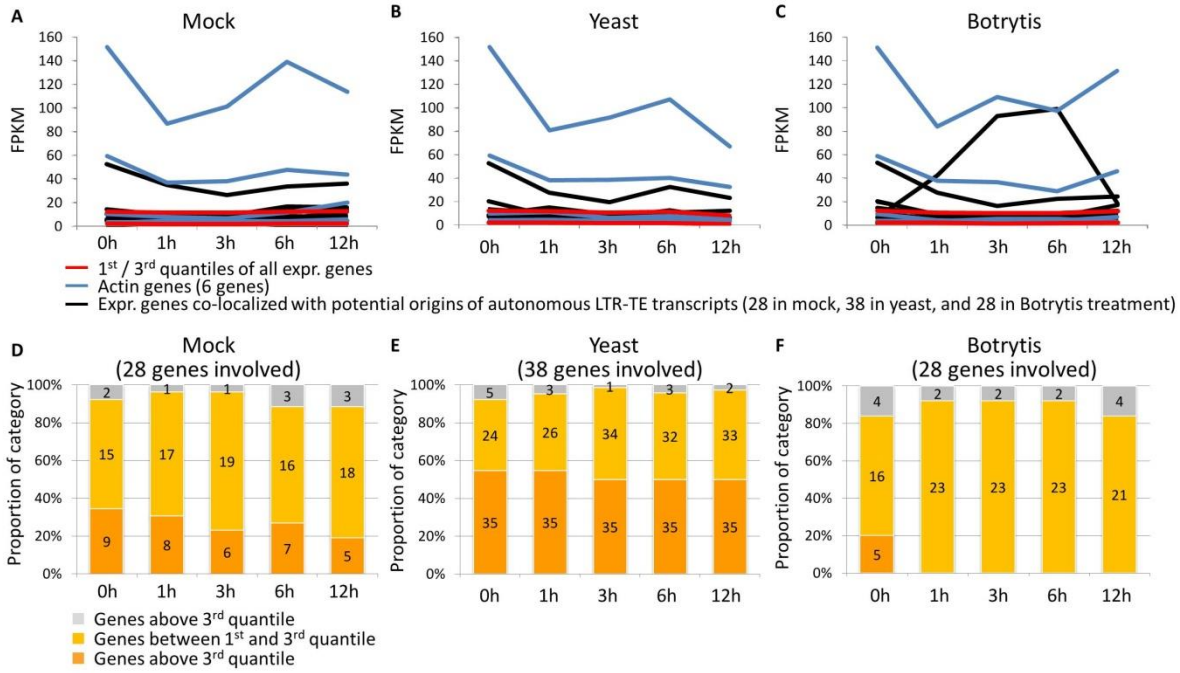



**Figure 3.16 Expression level of genes co-localized with the autonomous expression candidates**

(A-C) FPKM values of genes co-localized with the autonomous expression candidates were plot with those of genes encoding Actins. The first and third quantiles of all expressed genes were as indicated. (D-F) Genes associated with autonomous expression candidates were divided by the first and third quantiles.

The location bias towards introns is shown in Figure 3.14 and is particularly significant for both Copia-23 and Copia-3 autonomous expression candidates (Figure 3.15). Because Copia-23 and Copia-3 are considered to be the LTR-TE families that experienced the most recent transposition burst (Figure 3.11 and Table 3.2), we then questioned that, in the grapevine reference genome, whether the full-length autonomous loci of these two families were predominantly annotated in introns, and whether the full-length autonomous loci of other LTR-TE families exhibit location preference as similar as those of Copia-3 and Copia-23. To this aim, LTR-TE families were ordered in descent based on the number of structurally autonomous TE loci obtained in the families. From top 1 to top 5 are, respectively, Copia-23, Copia-31, Copia-3, Copia-33 and Gypsy-3 families (Figure 3.17 A). This order appears to be inversely related to the number of total elements in the respective families (Figure 3.17 B). Irrespective of the integrity, Copia-23 and Copia-3 elements were preferentially found in genic regions, whereas the other three families showed a tendency to be inserted in intergenic regions (Figure 3.17 C, D). Categorization of full-length TEs by the presence or absence of paired LTRs showed a similar proportion of autonomous over full-length loci among the five families (Figure 3.17 E). However, the predominance of genic insertions among autonomous TEs was only observed in Copia-23 and Copia-3 (Figure 3.17 F). A further level of location classification showed that the location bias towards introns appeared to be unique to these two most transcriptionally active LTR-TE families in *P. noir* embryogenic callus (Figure 3.17 G).

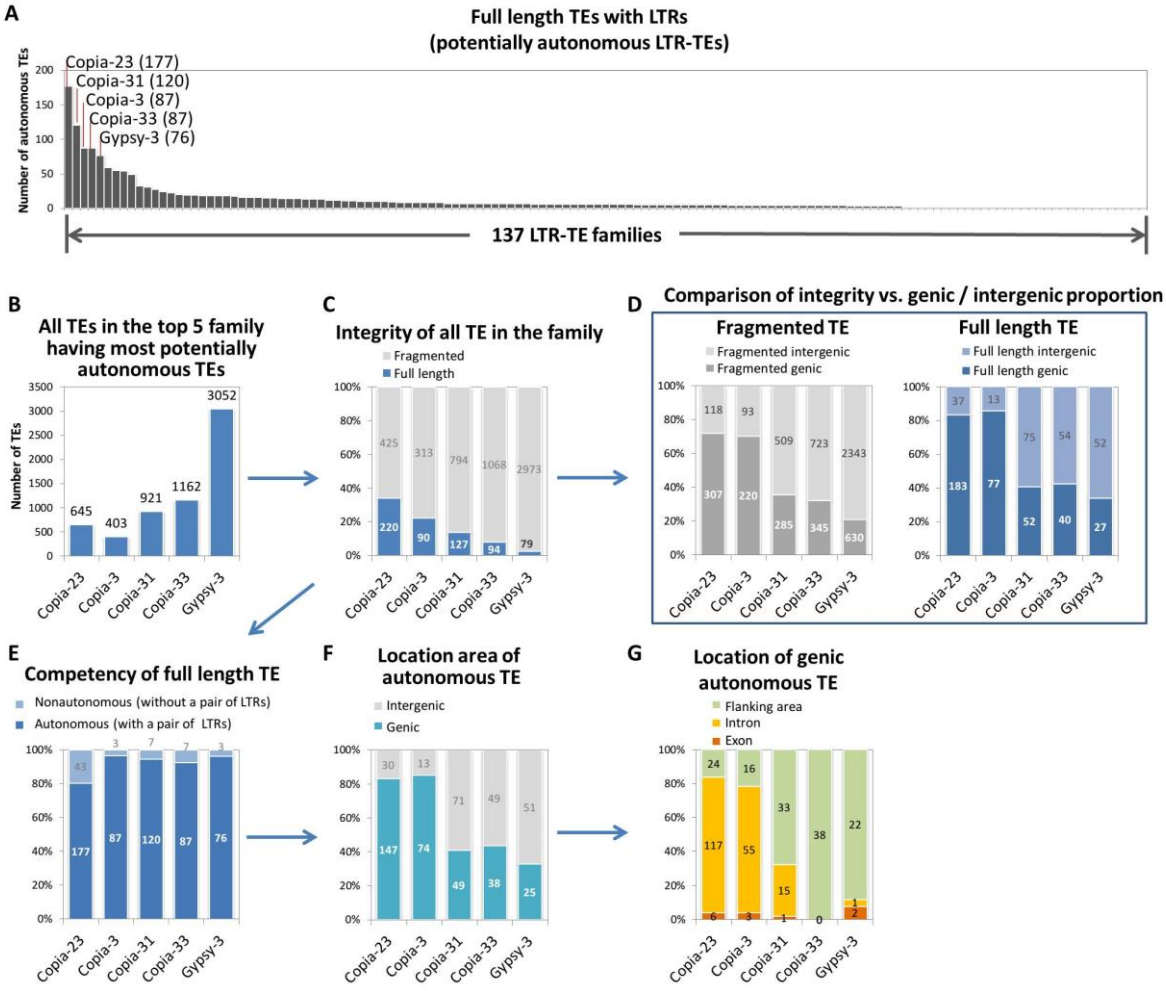



**Figure 3.17 Location of autonomous TEs from the top 5 families containing most autonomous loci.**

**(A)** 137 LTR-TE families ordered by the number of autonomous TEs (full-length TEs with pair of LTRs). The top 5 families were indicated. **(B)** The number of annotated TEs of the families indicated in (A). **(C)** The proportion of full length and fragmented TEs of the given families. **(D)** The location area of fragmented and full-length TEs. Copia-23 and Copia-3 have more genic loci than the other three families, irrespective of TE integrity. **(E)-(G)** Full-length TEs flanked with LTRs were extracted to analyse the location area (F), whereas genic autonomous TEs were categorized by location relative to genes (G). The number in or above each coloured area represents the number of TE loci of the corresponding category.

## 3.5 Discussion

### 3.5.1 Stress treatments increased transcriptional activity of TEs in terms of numbers of expression candidates

While there are large numbers of annotated TE loci (223,411 TE loci) in *V. vinifera* reference genome, less than 2% (3,698 TE loci) were identified as TE expression candidates in the embryogenic callus not exposed to any stress treatment (T=0), meaning that the majority of the TE loci are silenced in this tissue culture. The increase in the number of expression candidates from 3,698 TE loci in T=0 to 5,524 TE loci in mock treatment indicates that this treatment, supposedly resembling wound treatment, stimulated transcriptional expression of additional sets of TEs that were not seen in callus in the relatively steady state (T=0). Compared with mock, although the addition of the biotic stressors, yeast live cultures or *Botrytis* cell extracts, didn't substantially further increase the number of expression candidates, 2,298 TE loci were uniquely found to respond to the biotic stressors (yeast or *Botrytis* treatments) but not the wound treatment (Appendix C.2). Concordant with Lizamore (2013), these findings show that mock, yeast and *Botrytis* treatments are all able to induce the transcriptional activation of a group of TE loci that are silenced in the untreated embryogenic callus.

As mentioned in section 1.5, it is possible that these TE expression candidates contain wound- or pathogen-responsive *cis*-regulatory elements (CREs) that attract corresponding transcription factors to facilitate transcription initiation. The transcriptional activation of TE loci is frequently associated with relaxed epigenetic silencing on these loci, where a reduced level of DNA methylation or decreased abundance of H3K9me were observed (Dowen et al., 2012; Marí-Ordóñez et al., 2013; Rakocevic et al., 2009). These epigenetic changes might need to take place prior to the CRE-mediated stress response because, in *Arabidopsis*, impaired RdDM pathway (e.g. Pol IV mutant *nrpd1* and Pol V mutant *nrpd2*) was found to be the prerequisite of heat-induced transposition of TEs (*ONSEN*) that contain heat-responsive CREs (Ito et al., 2011). That being said, in our case, even if a TE locus does contain the CRE responding to the establishment of embryogenic callus, wounding or pathogens, this locus might not be able to be transcribed without the relaxation of epigenetic suppression. This chapter hence focuses on the factors (discussed as follows) that might affect the tendency of a TE locus to be targeted by the epigenetic silencing system.

In addition to the possession of stress-responsive CREs, the TE expression candidates identified in T=0, mock, yeast and *Botrytis* treatments might also be located in epigenetically relaxed genomic regions that house important genes for viability or stress response. Having been introduced in chapter 1 (section 1.4 and section 1.6.3), TE insertions within or close to important genes can cause a dilemma for host cells as intensive epigenetic suppression on these sites might also inhibit gene activity. To maintain the expression level of genes co-localized with these TEs, the epigenetic

suppression on these TE loci might be compromised or tuned down to the level that allows minimum host gene activity. In our study, therefore, it is plausible that, regardless of the presence or absence of CREs, TE expression candidates positioned in these transcriptionally permissive areas are able to take advantage of the epigenetically relaxed status and thus have a higher stochastic chance to access the transcription machinery (e.g. RNA Pol II) of host cells.

There are other factors that might also be the determinants of the tendency of a TE locus to be targeted by epigenetic silencing. The epigenetic machinery tends to target 'high-risk' TE loci that are capable of autonomous transposition. These TE loci are typically full-length and comprise sequence compartments representing intact key structural components (e.g. LTR or TIR) or encoding functional proteins (e.g. RT or TPase) necessary for autonomous mobilization (Panda et al., 2016). In other words, TE loci that are fragmented and have lost vital structural components or have accumulated sequence polymorphisms (e.g. SNVs and INDELs) and thus are incapable of autonomous mobilization are proposed to be more likely to be transcribed than the full-length autonomous TE loci.

Therefore, how the transcriptional activity of TE expression candidates correlated with these factors, including TE integrity, sequence polymorphisms (corresponding to the trackable and un-trackable expression candidates), TE location and TE sequence structure, are discussed in the following sections.

### 3.5.2 Transcribed TE loci are mostly fragmented and trackable by sequence polymorphism

Comparisons of the length of each element at any given locus and the corresponding canonical sequence reveal that over 95% of the annotated TEs are at least 10% shorter than the canonical element (Figure 3.4 A). In the embryogenic callus not exposed to any stress treatment (T=0), about 91% of the 3,698 expression candidates are fragmented (Figure 3.4 B). For embryogenic callus that was exposed to mock, yeast and *Botrytis* treatments, fragmented expression candidates contributed to a similar proportion (90-91%) of the total expression candidate pools. These results are concordant with the observations of Panda et al. (2016) in *Arabidopsis*, in which small TE loci (< 2kb) are lacking CHH methylation and less likely to be targeted by the expression-dependent form of epigenetic silencing (involving Pol II, RDR6 and Pol V; see section 1.3.2), whereas large and structurally intact TE loci are enriched with CHH methylation and tend to be targeted by the expression-dependent epigenetic silencing pathway. The findings in *Arabidopsis* mean that, compared to long autonomous TE loci, short and fragmented TE loci are more frequently associated with relaxed epigenetic silencing level in their sequence context, and they are more likely to escape

from the epigenetic re-suppression following after transcriptional activation. This might explain the extreme proportion of fragmented TE loci in the expression candidate pool.

Note that the maintenance of canonical RdDM (section 1.3.1) and the expansion of PTGS signal (section 1.3.2) rely on the sequence homology between siRNAs and targeted TE loci. If a TE locus has accumulated too many SNVs and INDELs to be recognized by siRNAs derived from highly conserved TE sequences, this TE locus might escape from the epigenetic silencing pathway and thus have a higher chance of transcriptional activation than the highly conserved TE loci. In grapevine embryogenic callus of all experimental conditions, over 70% of the TE expression candidates are trackable by unique-mapping reads (Figure 2.9), meaning that these expressed TE loci have obtained unique polymorphisms to be distinguished from other conserved TE sequences. This agrees with the assumption mentioned in section 3.5.1.

Combined analysis of the integrity and uniqueness (trackable or un-trackable) of expression candidates show that the majority of these are fragmented and trackable loci (Figure 3.5 – Figure 3.8), which are very likely old TE insertions of ancient TE families and are incapable of autonomous mobilization. From the opposite perspective, if a TE family is enriched with full-length un-trackable expression candidates, this family might be mobile recently, in the evolutionary time, and contribute new TE insertions due to the current transcriptional activity. This assumption is discussed below.

### 3.5.3 LTR-TE families showing most full-length untrackable expression candidates are likely to contribute competent transcripts for mobilization

As shown in Figure 2.9, the expression candidates can be grouped into trackable or un-trackable categories depending on the presence or absence of unique-mapping reads. The number of un-trackable candidates of any given TE family can be an indicator of the family's ability to mobilise autonomous and non-autonomous TE family members.

The dynamics of a TE family's mobilization varies across evolutionary time. In general, the overall transposition activity peaks at an early stage of TEs presence in the genome and then gradually decreases due, in large part, to the epigenetic inhibition and the consequential accumulation of sequence polymorphism, which would eventually render these members of the TE family inactive (Huang et al., 2012). As members of the element family age further, they decay in length, which can be resulted from large deletion or the self-recombination of identical repeats within any TE locus, and accumulate an increasing proportion of SNVs and INDELs until the elements are largely unrecognizable in the genome. Two assumptions can be derived from this concept. Firstly, TE families that have mobilized recently are likely to be recognized due to the identical or highly similar sequence at individual loci. In other words, attributing RNA-seq reads to individual loci will be

impossible due to the unavoidable multi-mapping of short sequencing reads. Secondly, TE families with more numbers of identical and structurally intact TE loci are more are likely to be able to contribute new insertions than other families.

Among all TE families with expression candidates, four LINE families contributed a considerable number of expression candidates (> 100 expression candidates per family) at T=0 (Figure 3.5 C). In mock, yeast and Botrytis treatments, the number of LINE families contributing more than 100 expression candidates in each family rose to nine from the ten families present in the genome (Figure 3.6 C, Figure 3.7 C, and Figure 3.8 C). However, most of these LINE expression candidates are fragmented and have accumulated polymorphisms (i.e. trackable) that render these loci to be individually distinguished by unique-mapping RNA-seq reads (i.e. fragmented and trackable loci; Figure 3.5 D, Figure 3.6 D, Figure 3.7 D, Figure 3.8 D). In fact, among all 23,447 annotated LINE loci in the reference genome, 23,268 (99.2%) are fragmented elements. These together suggest that the majority of LINE loci in the grapevine are relics of ancient LINE insertions, and they have been largely eroded (i.e. fragmented and diverged by mutations) through evolutionary time. These diverse polymorphisms in LINE expression candidates might facilitate their escape from the epigenetic silencing pathway that requires certain homology between siRNAs and the target loci, and therefore result in the transcriptional permissive state of these LINE loci. Despite the noticeable transcriptional activity of these degenerated LINE loci, the transcripts derived from these loci are unlikely to contribute to autonomous mobilization.

As opposed to LINEs, Copia-3 and Copia-23 stood up as the two LTR-TE families having the most full-length untrackable expression candidates (Figure 3.5 D). Given that the number of full-length candidates and the number of families having these full-length candidates increased in mock, yeast and *Botrytis* treatments, Copia-3 and Copia-23 are still the only two families each having more than 50 full-length untrackable candidate loci (Figure 3.6 – 3.8), suggesting that these loci are relatively young TE insertions and have not degenerated into short fragments or accumulated a large number of polymorphisms. Therefore these Copia-3 and Copia-23 full-length un-trackable expression candidates are more likely to contribute full-length transcripts competent in producing new insertions autonomously.

The abovementioned assumption is split into two parts to facilitate the examination of it. Firstly, the total sequence divergence of all full-length annotated loci of Copia-3 and Copia-23 was analysed and presented as neighbour-joining trees to test whether there exist clusters representing the five categories of full-length TE loci (section 3.3.3). These categories, ordered by the possibility to contribute autonomous mobilization, are structurally autonomous (full-length and flanked by LTRs) untrackable expression candidates, structurally autonomous trackable expression candidates,

structurally non-autonomous (full-length but lost at least one LTR) un-trackable expression candidates, structurally non-autonomous trackable expression candidates, and non-expressed full-length loci. It is expected that TE loci that were identified as structurally autonomous untrackable expression candidates would be clustered into a dense broom-like branchlet that exhibits high similarity shared by the sequences decent from the same branch. This form of branchlet is usually indicative of a recent proliferation of the TE family (Tsukahara et al., 2009). In the second part of the examination for the aforementioned assumption, the divergence of a pair of LTRs was used as leverage to estimate the insertion date of the corresponding LTR-TE locus (section 3.3.6) because it has been proposed that the 5' and 3' LTRs evolve independently and diverse from each other in sequence context from the time the mobilization of this single LTR-TE element took place (Moisy et al., 2008; SanMiguel et al., 1998). It is expected that the latest mobilization burst estimated from the LTR divergence of structurally autonomous Copia-3 and Copia-23 loci is more recent than that of other LTR-TE families.

The neighbour-joining tree of Copia-3 sequences from 90 annotated full-length loci revealed a condensed cluster containing 19 of the 20 sequences from un-trackable loci retaining LTRs (i.e. structurally autonomous un-trackable loci; Figure 3.9). Their un-trackable and clustering characteristics indicate their high identity, implicating the possibility of a recent burst giving rise to these highly similar insertions, as well as the potential for new autonomous movement. And if there was any production of autonomous Copia-3 transcripts in response to the stresses, it should be produced from some or one of them. On the other hand, the cladogram of 220 annotated full-length loci of Copia-23 showed that the 11 structurally autonomous un-trackable loci were separated into five clusters (Figure 3.10), suggesting that there has been a larger divergence of Copia-23 compared to Copia-3.

Despite the richness of full-length un-trackable candidates found in Copia-3 and Copia-23, it is possible that these candidates were picked up by the pipeline due to the identical sequences shared with fragmented un-trackable loci where the transcription was actually taking place. However, this possibility has been disproved by the analysis that categorized all sequencing reads (RNA-seq) associated with Copia-3 or Copia-23 into the four categories according to their mapping destinations: un-trackable fragmented, un-trackable full-length, trackable fragmented and trackable full-length (Appendix C.4). If the abovementioned scenario existed, all the Copia-3 and Copia-23-derived sequencing reads would be included completely in the un-trackable fragmented category, and no reads would be uniquely grouped into the un-trackable full-length category as a stand-alone subset in the Venn diagrams in Appendix C.4. In fact, none of these categories alone can fully include reads of one another category, meaning that the fragmented un-trackable expression candidates didn't

produce all the reads mappable to full-length un-trackable expression candidates and *vice versa*, and thus there must be one or more full-length un-trackable loci contributed to the sequencing reads mapping to expression candidates of their own category.

As previously mentioned, the second part of examining the activity of Copia-3 and Copia-23 is to further check whether Copia-3 and Copia-23 transposed more recent than other LTR-TE families by estimating their peak insertion time inferred from the polymorphisms of LTR sequences (see section 3.3.6). This analysis demonstrated that, among the 41 LTR-TE families each obtaining more than ten full-length loci with LTR pairs (termed complete copies or structurally autonomous loci), Copia-3 and Copia-23 are the families having most recent insertion peaks (Table 3.2), and the insertion date distribution of their complete copies was significantly more recent than other expressed LTR-TE families having full-length expression candidates (Figure 3.11). These observations strengthen the possibility that Copia-3 and Copia-23, the two LTR-TE families that experienced the most recent burst about 19 and 166 thousand years ago, respectively, might be capable of producing autonomous transcripts in response to stresses.

### 3.5.4 Expression candidates and the potential origins of autonomous transcripts tend to be found in introns of expressed genes

Given that TEs locate in different parts of genomes (e.g. exons, introns, promoters, and intergenic regions) can cause different degrees of outcomes on gene expression, they may be targeted by various degree of epigenetic suppression to maintain the expression of essential genes (section 3.2.2.2). In other words, the transcriptional activity of TEs might be correlated with their location relative to genes. Indeed, TE location analysis (section 3.4.3) shows the strong location bias of expression candidates towards expressed genes. In the grapevine reference genome, about half of annotated TEs were in exon, intron, or the 2kb-flanking regions of genes (Figure 3.12 B). However, in untreated embryogenic callus (T=0), over 70% of the expression candidates are positioned in genic regions, particularly introns of expressed genes (Figure 3.12 C). With the presence of stressors (mock, yeast and *Botrytis* treatments), the proportions of genic expression candidates were further increased by 5 per cent (Figure 3.12 D-F).

To facilitate the deduction of possible reasons underlining the distribution bias of TE expression candidates, the bias toward genic location is firstly discussed as follows, from which extended to the discussion of bias toward introns.

The genic location preference of expression candidates may be explained by two rules proposed as follows. Firstly, TE families that exhibit different insertion preferences (genic versus intergenic) might have different transcriptional activity in terms of the number of expression candidates. In this rule,

TE families predominantly distributed in the genic region contribute more expressed TE loci to the pool of expression candidates than families preferentially annotated in the intergenic region. Secondly, irrespective of the various genomic distribution patterns of different TE families, TE loci close to genes are more likely to be expressed. In other words, both TE families favouring genic and intergenic regions tend to contribute their genic TE loci as expression candidates. If this is the case, it is expected to observe a higher genic proportion in expression candidates grouped by families than the default genic proportion preset in the original distribution of all annotated TE loci of corresponding TE Families. Note that these two rules are not mutually exclusive. It is possible that some TE families demonstrate both phenomena.

The categorization of annotated TE insertions and expression candidates by families and genic/intergenic regions shows that LINE families that largely contributed expression candidates (although mostly fragmented and trackable; Figure 3.5-Figure 3.8) are TE families that preferentially annotated in genic regions (Appendix C.5 Figure C.7 A), hence supporting the first assumption. In addition, eight of the ten LINE families show that the genic proportion of LINE expression candidates grouped by families (Appendix C.5 Figure C.7 B-E) was further increased from the genic proportion of annotated LINE loci (Appendix C.5 Figure C.7 A), suggesting that the second assumption is also possible. In fact, the increase of genic proportions in expression candidates was generally seen in most of the TE families of Copia, Gypsy, hAT, and MULE (Appendix C.5). These observations support the second assumption and suggest that locating in the genic region might be a prerequisite for TE activation.

The same logic and analysis method was then utilised to compare the intronic proportion of all annotated genic TE loci and expression candidates to investigate the bias towards introns. This generated results (Appendix C.6) similar to the aforementioned findings of genic/intergenic distribution:

TE families (e.g. LINE families) that contributed more expressed loci to the pool of expression candidates are predominantly annotated in introns (Appendix C.6 Figure C.12 A), while other TE families (e.g. MULE families) that contributed less TE loci as expression candidates are not preferentially distributed in introns (Appendix C.6 Figure C.14 A). This variation in the transcriptional activity (in terms of the number of expression candidates) of TE families that favour different distribution in the reference genome can partially explain the overrepresentation of intronic expression candidates.

For most of the TE families (including LINE families), the intronic proportion of expression candidates is higher than the default intronic proportion estimated from the original distribution of annotated

loci (Appendix C.6), suggesting that, in general, intronic TE loci have a higher chance of transcriptional activation than other genic TE loci. This rule can be observed from TE families that preferentially distributed in introns (e.g. LINE families) and those not favouring intronic distribution (e.g. MULE families).

The observations from LINE families have shown that these two rules are not mutually exclusive, but the second rule (i.e. positioning in genic or intronic regions is a prerequisite of TE transcriptional activation) seems to be more determinative for TE transcription than the first rule (i.e. TE families that preferentially locate in genic or intronic regions are more transcriptionally active) since the second rule can benefit TE families that fulfil the first rule.

Although LINE families are set as good examples of these two rules, the majority of the LINE expression candidates are fragmented and largely diverse (i.e. trackable), hence are unlikely to mobilise autonomously. There are other two TE families, Copia-3 and Copia-23, that fulfil both rules (Appendix C.6 Figure C.10) and potentially seed autonomous mobilization (sections 3.4.4 and 3.5.3).

These two families are likely to be the LTR-TE families that experienced the most recent transposition burst (Figure 3.11 and Table 3.2), which might explain their low number of total annotated loci (Figure 3.17 B) but the high proportion of full-length loci (Figure 3.17 C) relative to the older LTR-TE families. TE loci of Copia-3 and Copia-23 are preferentially annotated in introns (Appendix C.6 Figure C.10 A), while the landscape of their expression candidates further enhances the intronic distribution bias (Appendix C.6 Figure C.10 B-E). Their recent mobilization burst, relatively low number of annotated loci, and the intronic tendency of the annotated and transcriptional landscape fit with a proposed theory, which suggests a mobilization cycle that might positively reinforce intronic insertion that predetermines the transcriptional and thus transpositional activity. This theory is discussed as follows.

It has been proposed that TE families with low copy numbers (less than a few hundred per genomes) prefer integrating into the genetically active part of the genome, gaining the opportunity for transcription and mobilization (Bennetzen, 2000). Despite the low-copy-number TE families could not expand in the genome efficiently by harnessing the limited number of TE loci, they might alternatively use genetically active areas as leverage for securing transcriptional activity and therefore increasing mobilization efficiency. In maize, high-copy-number LTR-TE families tend to aggregate and nest within each other in intergenic regions, and form substantial portions of centromeres, telomeres, and heterochromatic blobs (SanMiguel et al., 1996). In contrast, some low-copy-number maize DNA transposons predominantly target genes (Cresse et al., 1995; Liu et al., 2009). In four rice strains, DNA transposon *mPing* proliferated from roughly 50 to over 1000 copies,

with enrichment in euchromatic areas having high gene density (Naito et al., 2006, 2009). The rice endogenous LTR-TE *Tos17*, with copy number as low as one to five, is inactive in normal conditions, but this family can mobilize through the establishment of tissue culture and becoming inactive again in regenerated plants (Hirochika et al., 1996). According to Miyao et al. (2003), active *Tos17* preferentially targets gene-rich regions over heterochromatin regions, with the new insertions three times more likely to be found in genic regions including exons and introns than in other regions. Their researches support the theory that low-copy-number TE families manage to target genic regions in order to acquire the advantage for transcription and transposition. Although in our case, the investigated TE loci have been in the grapevine genome for at least several thousand years, the intronic prevalence was retained in the two TE families, Copia-3 and Copia-23, that likely experienced the latest transposition burst and still exhibit potentially full-length transcriptional activity by the establishment of embryogenic callus, wounding, and biotic stresses. In concordant with Bennetzen‘s theory (Bennetzen, 2000), Copia-3 and Copia-23, respectively, obtain only 403 and 645 annotated loci, which are relatively low compared to other LTR-TE families that have established thousands of insertions in the genome. Moreover, Copia-3 and Copia-23 expression candidates were predominantly found in genic regions, particularly in introns of expressed genes, resembling the theory that young TE families take advantage of the genetically active part of the genome to increase their transcription possibility.

Although our findings are concordant with the abovementioned theory that associated with host gene activity, we cannot exclude the possibility that the transcription of genic TE expression candidates is initiated by surrounding active genes (Sigman and Slotkin, 2016) instead of the promoter of TE loci. This would generate aberrant transcripts that are chimaeras of gene and TE sequences and are vulnerable to nonsense-mediated mRNA degradation (He and Jacobson, 2015; Moore, 2005). Therefore it is believed that TE transcripts as part of a larger genic mRNA only contribute to a very small portion of the transcriptome, and the chances of mobilization coming from these transcripts is low. It is also plausible that the potentially complete transcription of structurally autonomous LTR-TE loci was partially contributed from fragmented loci that are identical to part of the sequence of the autonomous loci. These doubts highlight the limitation of high-throughput short-read sequencing and the demand for long-read sequencing technology.

Despite the aforementioned theory can link the location preference of TE expression candidates with the transcriptionally permissive genic region of the genome and suggests a “hitchhiking manner” of these TE loci, the presence of these TE insertions within or proximal to genes might dampen the transcriptional activity of the co-localized genes (Hirsch and Springer, 2017). Le et al. (2015) reported that, in *Arabidopsis*, the level of methylated CG and CHG on the intragenic TE loci is negatively

correlated to the expression level of host genes. In soybean (*Glycine max*) and common bean (*Phaseolus vulgaris*), Kim et al. (2015) found that the differential expression of paralogue genes between these two species is negatively correlated with differential non-CG methylation level and percentage coverage of TE sequences within the gene body of the paralogue genes. That being said, soybean paralogue genes comprised of more TE sequences in exons or introns tend to be subjected to a higher level of non-CG methylation and exhibit lower expression level than the common bean counterparts that generally contain fewer TE sequences in the gene body. Therefore, with the observations of Le et al. (2015) and Kim et al. (2015) as well as the findings in this chapter, it is tempting to speculate that the transcriptional activation of full-length TEs closing to genes, especially those in introns, could be a trade-off between co-silencing the host genes by strong epigenetic suppression and permitting low expression of these TE loci to secure the necessary level of gene transcription (section 1.6.3). In this scenario, genes containing full-length expression candidates are less likely to be highly expressed than genes containing fragmented expression candidates or genes without TEs. This possibility will be addressed in the next chapter.

In addition, the downside of this trade-off strategy is that low-level expression of intragenic TEs might increase the chance of the production of aberrant transcripts as hybrids of genes and TEs (Sigman and Slotkin, 2016), and hence adulterating the overall expression of host genes with incomplete transcripts that miss downstream exons or display premature polyadenylation (Saze et al., 2013). This possibility again emphasizes the requirement of long-read sequencing data to interrogate the integrity of gene transcripts in association with intragenic TEs.

## 3.6 Conclusions

For a better understanding of the factors that contribute to TE activity, grapevine's embryogenic callus treated with wounding, yeast live cultures or *Botrytis* cell extracts can be used as a platform to stimulate TE's transcriptional activity. The collected TE loci that are potentially expressed are denoted as expression candidates. While the majority of TE expression candidates are defective remnants incompetent in autonomous transposition, only a few hundreds of full-length expression candidates are likely to contribute new transposition. Among the 232 TE families in the genome, Copia-3 and Copia-23 obtain the most full-length expression candidates across all experimental conditions. It appears that these two families are the two LTR-TE families that are most likely to achieve autonomous mobilization since they are found to experience the most recent transposition burst and still sustain a significant number of structurally autonomous loci that are highly conserved in sequences and potentially show complete transcription in grapevine embryogenic callus.

Analysis of the genomic landscape of expression candidates reveals a significant location bias towards introns, suggesting that intragenic TEs, in general, possess a higher opportunity of transcriptional activation than intergenic TEs. The over-representation of expression candidates, especially structurally autonomous LTR-TE candidates, within introns of expressed genes further supports the assumption that genetically active regions might shelter TE loci from extreme epigenetic silencing and provide the opportunity for transcription and mobilization. Nonetheless, permitting transcription of intragenic TE loci might risk the expression of host genes by attracting suppressive epigenetic silencing machinery to the surrounding regions or generating aberrant transcripts and exposing premature termination codons. Therefore the transcriptional dynamics of TEs and co-localized genes were further investigated in the next chapter.

# Chapter 4

# Association between TE and gene expression

## 4.1 Overview

The previous chapter reports the factors that may be crucial to transcriptional activation of TEs by investigating the location and characteristics of potentially active TEs (we termed 'expression candidates') in grapevine embryogenic callus induced by mechanical and biotic stress. A considerable proportion of annotated genes were found co-localized with TEs. From the perspective of genes, their transcriptional activity might be influenced by the presence of TE insertions.

In this chapter, the comparison between genes co-localized with TEs and those without TEs shows that genes containing TEs in exons or introns tend to be associated with a lower transcriptional level than genes without TEs. As TEs within genes are a possible lightning rod for epigenetic silencing, the suppressive modifications on TE sequences or surrounding chromatin structure may hinder the host genes from achieving higher transcriptional activity. On the other hand, the integrity of TE-expression candidates appeared to be negatively related to gene activity. Based on a currently developed model of TE silencing, the possible explanation is that the full-length TE-expression candidates were preferentially targeted by expression-dependent forms of RNA-directed DNA methylation (RdDM). This could further trigger conformational changes of chromatin that is permissive but prevent co-localized genes from achieving higher levels of transcription.

Leveraging the time-series experimental design used in earlier chapters, the dynamic changes of co-localized TEs and genes were investigated to understand their transcriptional behaviour with respect to both biotic and abiotic stimuli. The analysis revealed the tendency toward similar expression patterns from co-localized and differentially expressed TEs (DETEs) and genes (DEGs). Although it's not clear that whether there is a causal relationship between the activity of TEs and genes, it is plausible that chromatin context around a paired DETE and DEG were similar, therefore allowing synchronous transcriptional dynamics.

## 4.2 Introduction

As mentioned in chapter 1, TE insertions within genes or at the promoters may have a considerable impact on gene expression, either genetically or epigenetically, in terms of the degree of transcriptional initiation, inhibition, or alternative splicing. However, the outcomes differ between TE families, tissue types, and species.

In some species, like maize and Norway spruce, intragenic TEs generally exhibited little influence in the expression of host genes (Nystedt et al., 2013; West et al., 2014), whereas in tomato, in particular, the case of the tomato gene *VTE3*, which encodes an enzyme mediates vitamin E synthesis, showed different levels of gene transcripts associated with distinct TE insertions within the promoter region (Quadrana et al., 2014). Quadrana and colleagues (2014) also found that the *VTE3(1)* alleles were differentially expressed in *Solanum lycopersicum* and *S. pennellii*: the former exhibiting lower *VTE3(1)* activity than the latter. The *VTE3(1)* promoter regions in both tomato species contain two copies of the DNA transposon *Tc1-IS630-Pogo*, whereas a TE insertion of the SINE family was additionally discovered in the same region of *S. lycopersicum*. The presence of hypermethylation and accumulation of 24 nt siRNAs at the TE insertions in *S. lycopersicum* but not in *S. pennellii* suggest the differential *VTE3(1)* epiallele activity determined by the distinct TE insertion patterns at the promoter regions (Quadrana et al., 2014).

In *Arabidopsis thaliana*, Hollister and Gaut (2009) investigated the impact of intergenic TE insertions on the expression level of nearest genes and found that genes nearest to an intergenic TE insertion were significantly expressed at a lower level compared with the expression level of the rest of genes. In addition, the density of hypermethylated intergenic TEs within 10 kb distance to genes was found negatively correlated with the expression level of nearest genes (Hollister and Gaut, 2009). While Hollister and Gaut (2009) excluded genes containing intragenic TEs from their analysis to focus on the influence of intergenic TEs on nearest genes, Le et al. (2015) interrogated the relationship between intragenic TEs and the transcriptional level of host genes in the same species. They categorised *Arabidopsis* genes by the presence or absence of intragenic TEs, and further grouped genes containing TEs by the CHG methylation level of corresponding intragenic TEs. They found that genes containing TEs tended to have lower expression level than genes without TEs, and the expression level of genes housing TEs was inversely related to the CHG methylation level of TEs (Le et al., 2015). The comparison between soybean (*Glycine max*) and common bean (*Phaseolus vulgaris*) shows soybean genes that contain more TE sequences in gene body than the corresponding paralogue genes in common bean tend to have lower transcription level and higher non-CG methylation level than their paralogue counterparts in common bean generally (Kim et al., 2015; see section 3.5.4 for more details).

Up-regulated gene expression associated with TE insertions was also reported. For instance, a MER41 TE was found serving as an enhancer of the human *AIM2* (absent in melanoma 2) gene participating in inflammation response (Chuong et al., 2016). As previously mentioned, it has been reported that the peppered moth adapted to air pollution at the height of the industrial revolution and manifested itself by increasing numbers of darker individuals in the population. This transformation of the population was associated with an intronic TE insertion that enhanced the expression of the gene *cortex* through a yet unknown mechanism (Van't Hof et al., 2016).

Overall, most intragenic TE insertions (particularly those in introns) that have been retained through evolutionary time seem to have neutral or mild negative effects on gene expression. In comparison, deleterious insertions can be expected to have been removed from the population due to purifying selection. A recent study harnessed *Arabidopsis thaliana* epigenetic recombinant inbred lines (epiRILs) to exponentially generate TE insertions through 16 generations. It showed that new TE insertions were preferentially found near or within genes and that the epigenetic response evolved *in situ* to restore the expression of host genes (Quadrana et al., 2019). In their study, two host genes with new intronic *ATCOPIA93* (*EVD*) insertions showed a tendency of the reduced transcriptional level at F8 but regained expression activity at F16, when hypermethylation had been established on new copies of *EVD* (Quadrana et al., 2019). However, it is not clear that whether the expression activity of these two genes in F16 is restored to the level comparable to the original (wild-type) transcription level since the relative expression level of these genes between F16 and wild-type was not presented in Quadrana et al. (2019).

Based on the findings of Quadrana et al. (2019), it is plausible that the establishment of the cytosine methylation on intragenic TEs is pivotal to maintain the transcriptional activity of host genes. However, these host genes might not be able to be highly expressed if the negative correlation of TE methylation level and host gene transcriptional level represented by Le et al. (2015) and Kim et al. (2015) is also taken into consideration. That being said, the established DNA methylation in the gene body as a consequence of intragenic TE insertion might facilitate the restoration of the transcriptional activity of host genes but likely forbid these genes to be highly expressed. Because it has been proposed that full-length TEs are preferentially targeted by the expression-dependent form of epigenetic silencing (see section 1.3.2; Panda et al. 2016), and that the epigenetic regulation of TEs and their roles in gene regulation are specified by the location of TEs in the genome (section 1.4; Sigman and Slotkin, 2016), we questioned that whether the relatively low expression activity of genes containing TEs in Le et al. (2015) and Kim et al. (2015) could be observed in grapevine embryogenic callus and whether this phenomenon was possibly associated with the transcriptional activity, integrity, and position of TEs co-localized with genes.

The strong location bias of the TE expression candidates in grapevine embryogenic callus (see section 3.4.3) hints at the tolerance of intronic TEs to allow a certain level of gene transcription (see section 3.5.4). With the aforementioned examples included in chapter 1 and this chapter, two assumptions were raised. Firstly, it is possible that the transcriptional level of genes co-localizing with TEs would negatively associate with TE transcriptional activity, TE integrity, and intragenic insertion. Secondly, although gene activity is thought to be compromised by co-localized TEs, the over-representation of expression candidates in introns of expressed genes give rise to an assumption that intragenic TEs take advantage of the permissive transcriptional status (section 3.5.4) and therefore might display expression dynamics that resemble to host gene's activity.

To address the first assumption, we followed the analysis method of Le et al. (2015), in which the transcriptional level of expressed genes co-localized with TEs was compared with that of expressed genes not co-localized with TEs (hereinafter "genes without TEs"). For genes co-localized with TEs, we included TE transcriptional activity (i.e. expression candidates versus non-expression candidates), TE integrity (i.e. full-length versus fragmented) and TE location as potential variables of gene expression in our analysis (please see Methods section 4.3.1 for details). It is expected that expressed genes co-localized with full-length and transcriptionally active TEs would tend to have lower transcriptional level than expressed genes not co-localized with TEs. Although the expression level of genes co-localized with TEs might also be related to other factors, such as gene function and the presence of specific TE components (e.g. PBS and *gag*; see TE structure in chapter 1) that are preferentially targeted by small RNAs (Marí-Ordóñez et al., 2013; Panda et al., 2016; Schorn et al., 2017), these factors are likely all interwoven with the aforementioned factors including TE integrity, location and transcriptional activity. Because of the great variances of gene function and TE structural components, these two factors are better interrogated case by case and hence not included in this genome-wide analysis.

To interrogate the second assumption, i.e. TEs co-localized with expressed genes take advantage of the permissive transcriptional status and hence the concordant correlation of expression dynamics between TEs and host genes, the expression pattern of differentially expressed genes and co-localized transcribed TEs was examined. This analysis is only applicable for TE expression candidates obtaining unique-mapping reads (i.e. trackable expression candidates identified in chapter 2; see sections 2.4.2, 2.5.5 and 4.3.2 for details). We expected that the expression pattern of trackable expression candidates would be generally similar to that of the corresponding co-localized genes.

## 4.3 Methods

### 4.3.1 Comparison of the expression level of genes co-localized with TEs and genes without TEs

To minimise factors contributed from stress treatments, this analysis was only conducted on the data of untreated embryogenic callus (T=0) that was not subjected to mock, yeast, or *Botrytis* treatments (chapters 2 and 3). Using the RNA-seq data in chapter 2 and chapter 3, all annotated genes were categorized by their transcriptional activity in T=0 based on a cut-off threshold of FPKM > 1. Each group of genes (i.e. expressed and non-expressed genes) were further categorised hierarchically as follows:

(1) categorised by the presence and absence of co-localized TEs (i.e. whether the genes were with TEs or without TEs);

(2) categorised by the transcriptional activity of co-localized TEs (i.e. whether the co-localized TEs were expression candidates or non-expression candidates);

(3) categorised by the location of co-localized TEs (i.e. whether the genes contained TEs internally or were within 2kb of a TE locus);

(4) categorised by TE integrity (i.e. whether the genes co-localized with full-length TEs or fragmented TEs).

This information was used to plot Figure 4.1 using ggplot2 (Wickham, 2016).

For each category of expressed genes, the FPKM values of expressed genes were plotted into violin plot layered with dot plot and box plot indicating the quantiles and average FPKM value (Figure 4.2 – Figure 4.3). The categorization was processed using dplyr (Wickham et al., 2018) and illustrated using ggplot2 (Wickham, 2016). Statistical test on the mean of FPKM between categories was carried out using `t.test` in R.

### 4.3.2 Analysis of the expression pattern of TEs and genes

TE-expression candidates with unique-mapping reads (trackable TE-expression candidates) in at least one of the experimental conditions were collected for differential analysis of the expression dynamic changes under different treatments over time. For those captured by sub-pipeline 1 (Figure 3.2), their raw read count generated by htseq-count (Anders et al., 2015) was used for the analysis. The rest of the trackable expression candidates collected from sub-pipeline 3 (Figure 3.2) were analysed by using the raw read count of danglers produced by TEFingerprint. The analysis was conducted using

DESeq2 (Love et al., 2014) as read counts of the trackable candidates in all treatments (39 sequencing libraries) were analysed together with multifactor settings, where the full model of the design formula included treatments and time points and the reduced model included the factor of time only. Raw read counts of TEs with adjusted p-value under 0.05 were normalized and logarithmically transformed and normalized using the function `varianceStablizingTransformation` (VST) implemented in DESeq2 (Love et al., 2014). Afterwards, the VST-transformed read counts in mock treatment were normalized against that at T=0, while the VST-transformed read counts in yeast and *Botrytis* treatment were normalized with that of T=0, and then the effect from mock was further deducted at each time point. After these processes, TEs with the final normalized values above 1 in at least one time point were considered as differentially expressed TEs (DETEs). The Venn diagrams showing the unique and shared DETEs among the three stress treatments were generated using the R package VennDiagram (Chen and Boutros, 2011). Hierarchical clustering was conducted using `hclust` in R with the Pearson correlation method for measuring distances among DETEs and the "`complete`" agglomeration method for clustering. With the hierarchical clustering information, heatmaps of the final normalized value were plotted by `heatmap.2` from the package gplots (Warnes et al., 2020). For each DETE cluster, the expression trend revealed from the heatmap was illustrated with a simplified line graph. The statistics of the expression trend were summarized as a pie graph using ggplot2 (Wickham, 2016). Differential analysis and expression pattern clustering for genes were conducted in the same way as for TEs. Co-localized DETE and DEGs were gathered for hierarchical clustering as mentioned above to test whether the paired DETE and DEG would be grouped into the same expression cluster. Computational scripts used in this chapter can be found in Appendix D.3.

## 4.4 Results

### 4.4.1 Relationships between TE insertions and gene expression level at T=0

To understand the relationships between gene expression level and TE insertions in embryogenic callus without interference from abiotic and biotic stimuli, all annotated genes were firstly categorized hierarchically by transcriptional activity at T=0, the presence or absence of TE insertions, TE transcriptional activity, TE location, and the integrity of co-localized TEs (see Methods 4.3.1 for more details).

The exons of 31,845 annotated genes comprise 50,082,135 bases, representing 10.3% of the reference genome. These exons frame 116,128,035 bases as introns, which in turn make up 23.9% grapevine genome. With the establishment of embryogenic callus, the annotated genes were split in half by a cut-off threshold of FPKM=1 that produced 15,100 (47.42%) expressed and 16,745 (52.58%) non-expressed genes (the central core of Figure 4.1). Interrogation of both expressed and non-expressed genes revealed that 82.75% of the total annotated genes possessed co-localised TEs (dark green, 2$^{nd}$ layer of Figure 4.1). For the 12,612 transcribed genes co-localized with TEs, 1,955 (15.50%) of these genes were co-localized with TEs that were identified as expression candidates in T=0 (the upper red segment in Figure 4.1). Of this subset of genes, 1,559 (79.74% of the red segment) hosted TE-expression candidates within the gene unit (introns or exons; the top yellow section in Figure 4.1). The remaining 10,657 genes co-localized with TEs were associated with non-expressed TEs, and about half of these genes contained TE insertions in the gene unit (the second top yellow segment of Figure 4.1).

On the non-expressed half of annotated genes, 13,740 of the 16,745 unexpressed genes co-localize with TEs, and 422 of these genes were associated with TE expression candidates (the bottom red segment of Figure 4.1). One hundred and twenty-five of the 422 genes co-localized with TE-expression candidates housed these expression candidates in the gene unit (Figure 4.1, Table 4.1). Concordant with our previous observations, the majority of TE-related genes only hosted fragmented TEs (light blue segments in Figure 4.1). The number and percentage of genes of each category can be found in Table 4.1.

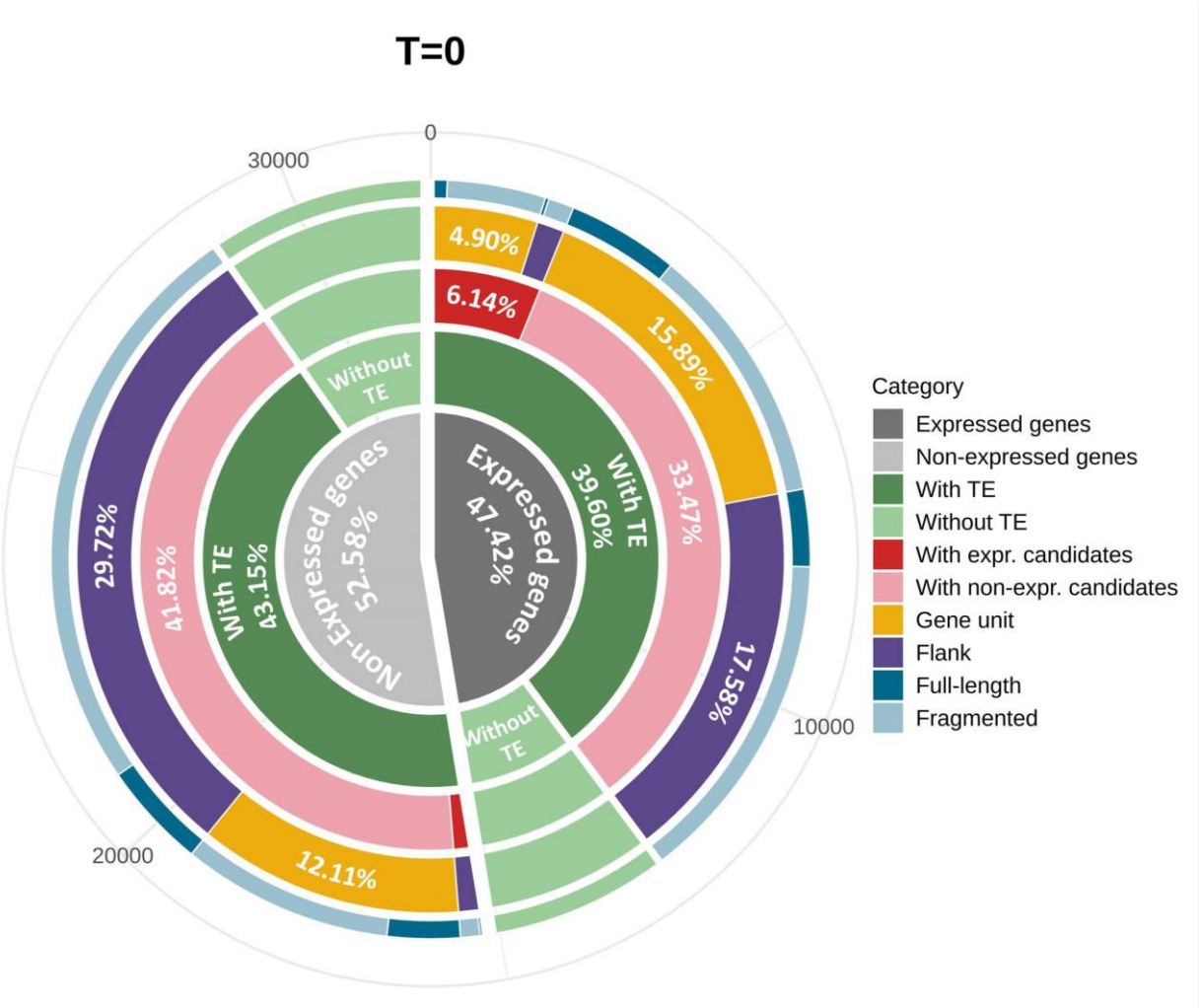



**Figure 4.1 Hierarchical categorization of all annotated genes**

From the inner-most layer of the graph, all annotated genes were categorized by gene activity, presence or absence of TE insertions, presence or absence of expression candidates, TE location, and TE integrity. Each category was denoted by specific colour as indicated.

**Table 4.1 Hierarchical categorization of all annotated genes by gene activity and TE insertions.**

Numbers of genes (#Gene) in black colour sum up to 31,845 annotated genes, while the corresponding percentages (Perc.) in black colour sum up to 100.00%. Subtotals of indicated categories were denoted in grey.

| Gene activity | TE insertion | TE activity | TE location | TE integrity | T=0 | |
|---|---|---|---|---|---|---|
| | | | | | #Gene | Perc. |
| **Expressed gene** | With TE | With expr. candidates | Gene unit | Full-length | 224 | 0.70% |
| | | | | Fragmented | 1,335 | 4.19% |
| | | | *Subtotal (Gene unit)* | | *1,559* | *4.90%* |
| | | | Flanks | Full-length | 49 | 0.15% |
| | | | | Fragmented | 347 | 1.09% |
| | | | *Subtotal (Flanks)* | | *396* | *1.24%* |
| | | *Subtotal (With expr. candidates)* | | | *1,955* | *6.14%* |
| | | With non-expr. candidates | Gene unit | Full-length | 1,517 | 4.76% |
| | | | | Fragmented | 3,542 | 11.12% |
| | | | *Subtotal (Gene unit)* | | *5,059* | *15.89%* |
| | | | Flanks | Full-length | 1,047 | 3.29% |
| | | | | Fragmented | 4,551 | 14.29% |
| | | | *Subtotal (Flanks)* | | *5,598* | *17.58%* |
| | | *Subtotal (With non-expr. candidates)* | | | *10,657* | *33.47%* |
| | *Subtotal (With TE)* | | | | *12,612* | *39.60%* |
| | Without TE | Without TE | Without TE | | 2,488 | 7.81% |
| | *Subtotal (Without TE)* | | | | *2,488* | *7.81%* |
| *Subtotal (Expressed gene)* | | | | | *15,100* | *47.42%* |
| **Non-expressed gene** | With TE | With expr. candidates | Gene unit | Full-length | 18 | 0.06% |
| | | | | Fragmented | 107 | 0.34% |
| | | | *Subtotal (Gene unit)* | | *125* | *0.39%* |
| | | | Flanks | Full-length | 33 | 0.10% |
| | | | | Fragmented | 264 | 0.83% |
| | | | *Subtotal (Flanks)* | | *297* | *0.93%* |
| | | *Subtotal (With expr. candidates)* | | | *422* | *1.33%* |
| | | With non-expr. candidates | Gene unit | Full-length | 988 | 3.10% |
| | | | | Fragmented | 2,867 | 9.00% |
| | | | *Subtotal (Gene unit)* | | *3,855* | *12.11%* |
| | | | Flanks | Full-length | 1,446 | 4.54% |
| | | | | Fragmented | 8,017 | 25.18% |
| | | | *Subtotal (Flanks)* | | *9,463* | *29.72%* |
| | | *Subtotal (With non-expr. candidates)* | | | *13,318* | *41.82%* |
| | *Subtotal (With TE)* | | | | *13,740* | *43.15%* |
| | Without TE | Without TE | Without TE | | 3,005 | 9.44% |
| | *Subtotal (Without TE)* | | | | *3,005* | *9.44%* |
| *Subtotal (Non-expressed gene)* | | | | | *16,745* | *52.58%* |
| **Sum** | | | | | 31,845 | 100.00% |

After the hierarchical categorization of genes, the expression levels of different categories of expressed genes were then compared to reveal the relationship between TE insertion and gene activity. All categories of expressed genes containing TEs were compared with expressed genes without TEs in terms of FPKM distribution (Figure 4.2). Note that 'genes without TE' refers to 'genes not co-localized with TE', which excludes genes containing TE insertion internally and genes within 2 kb distance of a TE insertion.

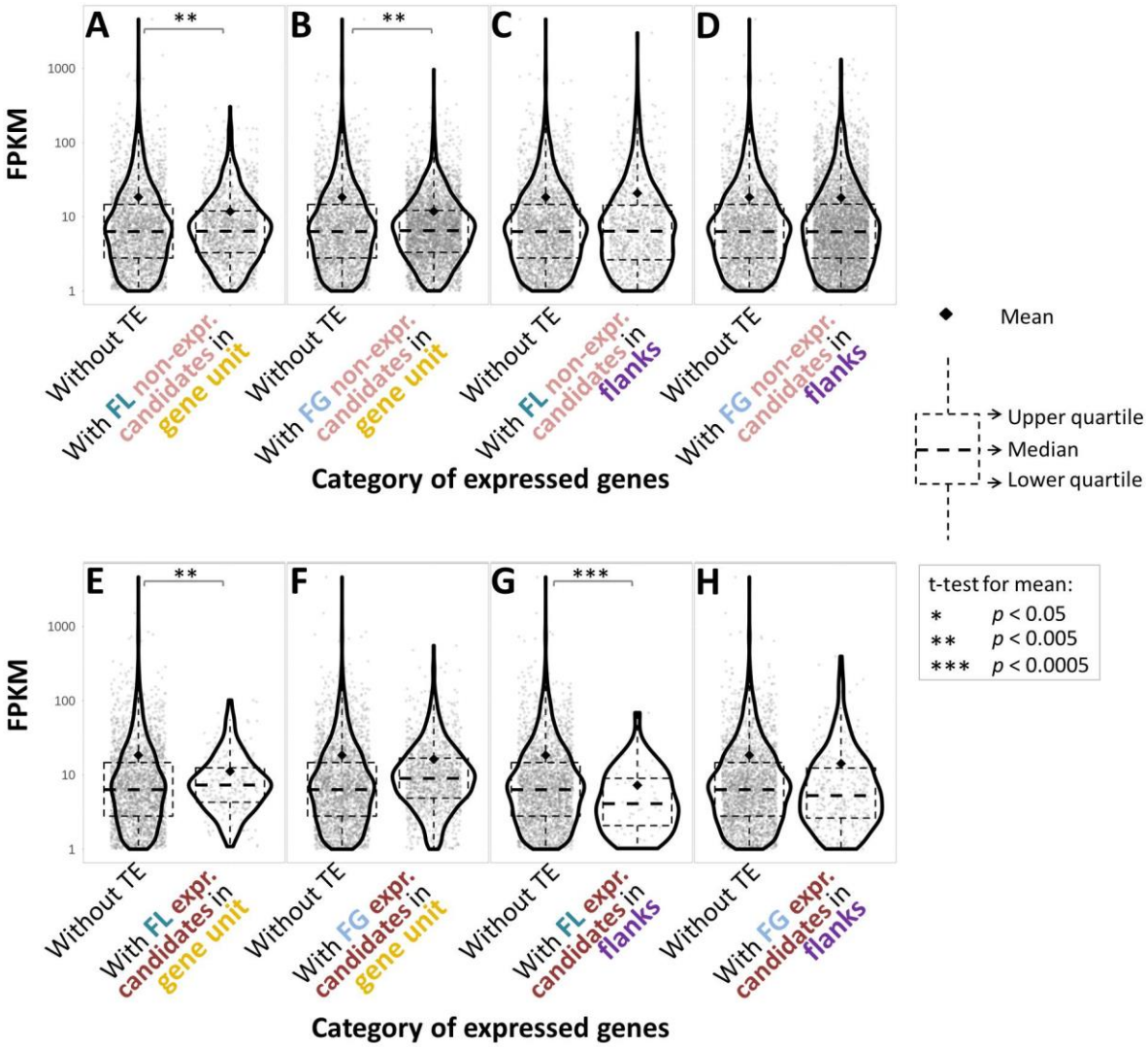



**Figure 4.2 Comparison of the expression level between genes without TE and those with TEs.**

The FPKM value of expressed genes not co-localized with TE (i.e. expressed genes 'without TE') was compared pair-wisely with expressed genes having **(A)** full-length expression candidates in gene unit, **(B)** fragmented expression candidates in gene unit, **(C)** full-length expression candidates in flanks, **(D)** fragmented expression candidates in flanks, **(E)** full-length non-expression candidates in gene-unit, **(F)** fragmented non-expression candidates in gene unit, **(G)** full-length non-expression candidates in flanks, and **(H)** fragmented non-expression candidates in flanks.

First of all, for genes containing intragenic TEs but none of these TEs was transcriptionally active (i.e. non-expression candidates), the expression level of these genes tended to be lower than that of genes without TEs, irrespective of the integrity of TEs within host genes (Figure 4.2 A, B). By contrast, if the unexpressed TE insertions were all in flanking regions, then no difference in gene expression was observed (Figure 4.2 C, D).

For the situation that transcriptionally active TEs (i.e. expression candidates) were present in the gene unit, those genes housing full-length expression candidates (Figure 4.2 E) exhibited expression level similar to the aforementioned situation for genes housing full-length unexpressed TEs within genes (Figure 4.2 A; p-value = 0.5627 for comparison of Figure 4.2 E versus Figure 4.2 A), and therefore genes housing full-length expression candidates in gene unit remained less transcriptionally active than genes without TEs (Figure 4.2 E). On the contrary, if the intragenic expression candidates were all fragmented, then the transcriptional level of these host genes was similar to that of genes without TEs (Figure 4.2 F).

Compared with the situation for genes within 2 kb distance to full-length unexpressed TEs (Figure 4.2 C), the presence of full-length expression candidates in the 2kb-flanking regions was significantly correlated with the lower expression level of co-localized genes (Figure 4.2 G; p-value < 0.0005 for comparison of Figure 4.2 G versus Figure 4.2 C), and thus these genes showed expression level significantly lower than genes without TEs. If the expression candidates that were present in the 2-kb flanking regions were all fragmented, then no difference in gene expression level was observed (Figure 4.2 H).

These observations suggest that if the co-localized TE loci were not expressed, the intragenic insertion of TE loci, irrespective of TE integrity, appears to be the variable that negatively correlated with the host gene expression level (Figure 4.2 A-D). If the co-localized TE insertions were transcriptionally active, the integrity of TE loci, regardless of TE location, seems to be the factor that negatively associated with the expression level of host genes (Figure 4.2 E-H).

### 4.4.2 Relationships between co-localized TEs and genes in terms of expression pattern across time

Although there exists a negative correlation between TE insertions and the expression of host genes (section 4.4.1), the over-representation of expression candidates in introns of expressed genes (section 3.4.3) indicates the tolerance of the transcriptional activity of intragenic TEs within expressed genes and suggests a "hitchhiker-like" manner of intragenic TEs in that they take advantage of the genic transcriptional permissive status for their own expression (as previously mentioned in section 3.5.4), and therefore these TEs might display expression dynamics that resemble a host gene's activity. Based on the observations in the previous section, genes co-localized with TEs may be less likely to be highly active among all expressed genes. However, these genes might retain the ability to be transcriptionally activated in response to stress in a relatively smaller range of expression level than genes without TEs. Additionally, their expression pattern over time might predetermine the transcriptional dynamics of the co-localized TEs.

To address this possibility, differential transcriptional changes of TE expression candidates and genes were detected by the computational tool DESeq2 (Love et al., 2014). Due to the repetitive characteristics of TEs, only a subset of expression candidates mapped by unique-mapping reads were suitable for this analysis (see sections 2.4.2 and 2.5.5). These expression candidates were termed "trackable expression candidates." The differential analysis was performed on 6,212 trackable expression candidates that were found in at least one of the four experimental conditions (T=0, mock, yeast and *Botrytis* treatments). This analysis revealed different sets of differentially expressed TEs (DETEs) in mock, yeast and *Botrytis* treatments (Figure 4.3). Hierarchical clustering of DETEs demonstrated various predominant expression patterns in response to different treatments (Figure 4.4). The mock treatment showed that roughly 50% of the DETEs were transcriptionally activated in the first 3 hours (3h) of post-treatment and then returned to an expression level similar to that observed at T=0 (Figure 4.4 A, B), illustrating an 'up-back' expression pattern. Figure 4.4 C shows that the most predominant expression change in mock treatment is the 'up-back' pattern. This pattern was also prevalent among DETEs of *Botrytis* treatment, yet with a tendency peaked at 6 hours (6h) of inoculation (Figure 4.4 G-I) as opposed to the prompt transcriptional activation at 1h and 3h in mock treatment. Interestingly, 206 of the 291 DETEs (70.79%) responded to *H. uvarum* incubation in an up-regulated manner (Figure 4.4 D-F).

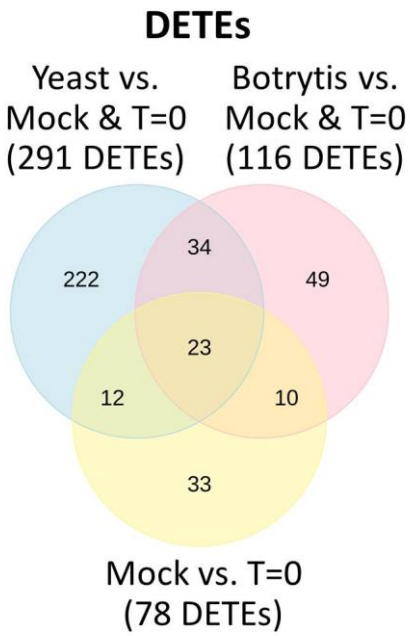



**Figure 4.3 Comparison of the three sets of DETEs responsive to mock, yeast and *Botrytis* treatment.**

DETEs of mock treatments denote TEs differentially responded to mock treatment against their initial state at T=0. DETEs of yeast or Botrytis treatment include those acting differentially from what they were at T=0 and mock condition. The overlapping areas demonstrate the shared number of DETEs, yet they may behave in different expression patterns over time in some of the treatments.

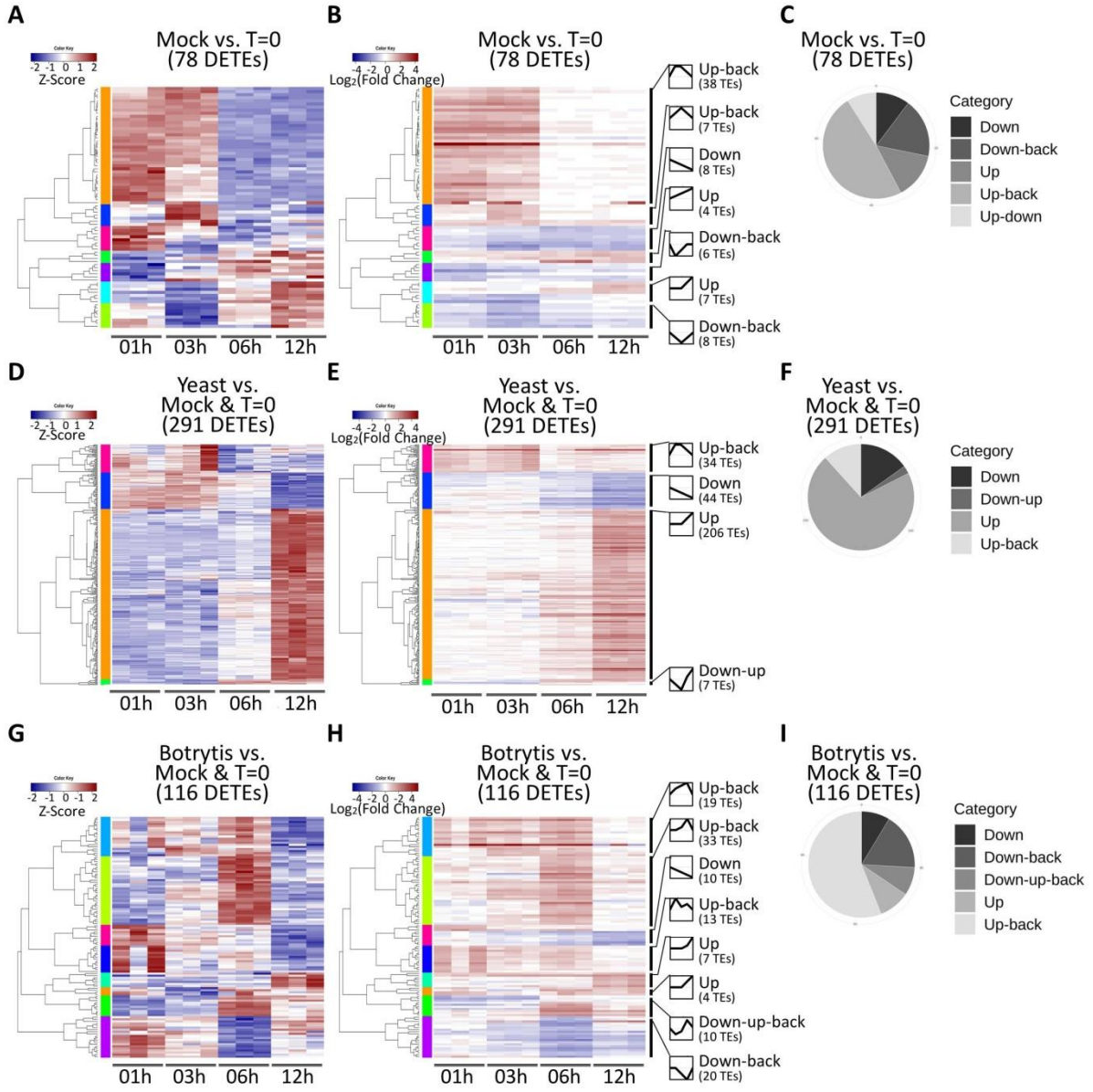



**Figure 4.4 Expression patterns of DETEs**

DETEs in mock, yeast and *Botrytis* treatments were illustrated by heat-maps using Z-score (A, D, G, respectively) or $\log_2$(fold change) (B, E, H, respectively), while the latter were labelled with line graphs representing the dynamics of each cluster over time. Clusters of similar expression patterns were then categorized together for the pie graphs summarizing the trend of expression changes in mock (C), yeast (F) and Botrytis (I) treatments.

Differential expression analysis on genes captured three sets of differentially expressed genes (DEGs), each responding to a specific treatment (Figure 4.5). The expression pattern of the DEGs in mock, yeast and *Botrytis* treatments (Figure 4.6) was similar to that observed for DETEs of the corresponding treatments (Figure 4.4). While most of the DEGs in mock treatment dropped back to T=0 basal levels after 3 hours of activation (Figure 4.6 A-C), the DEGs in *Botrytis* treatment tended to peak at 6h before dropping back to the initial state in 12h (Figure 4.6 G-I). Concordant with the most predominant pattern of DETEs with yeast treatment, 2,935 of the 5,060 DEGs (58.00%) were up-regulated through the latter half of yeast treatment (Figure4.6 D-F).

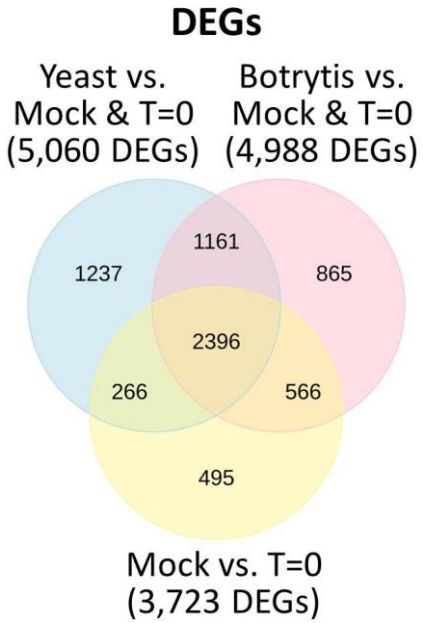



**Figure 4.5 Comparison of the three sets of DEGs responsive to mock, yeast and *Botrytis* treatments.**

DEGs of mock treatments are genes differentially responded to mock treatment against their initial status at T=0. DEGs of yeast or Botrytis treatment include those acting differentially from what they were at T=0 and mock condition. The overlapping areas demonstrate the shared number of DEGs, yet they may behave in different expression patterns over time in some of the treatments.

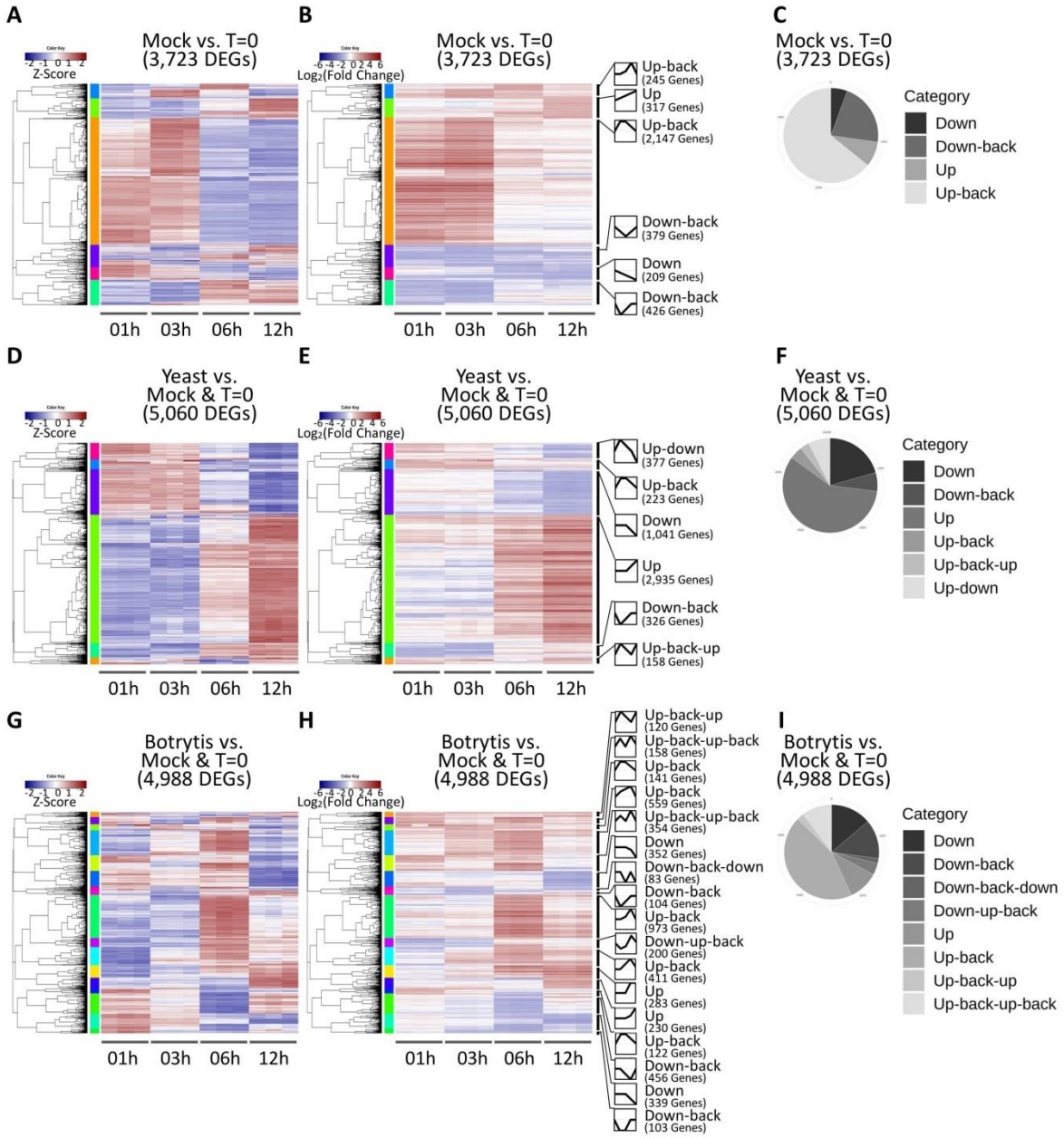



**Figure 4.6 Expression patterns of DEGs**

DEGs in mock, yeast and *Botrytis* treatments were illustrated by heatmaps using Z-score (A, D, G, respectively) and $\log_2$(fold change) (B, E, H, respectively), while the latter were labelled with line graphs representing the dynamics of each cluster across time. Similar clusters were then grouped together for the pie graphs summarizing the trend of expression changes in mock (C), yeast (F) and Botrytis (I) treatments.

In an attempt to investigate the relationship of expression pattern between DEGs and the co-localized DETEs, these corresponding DETEs and DEGs were used for hierarchical clustering, in which DETEs and DEGs of similar expression pattern were grouped into the same clusters. An initial survey of the DETEs under mock condition showed that 62 (79.5%) of the 78 DETEs were in the genic region, of which 4, 13, and 45 were respectively co-localized with unexpressed genes, expressed genes but not DEGs (i.e. non-DEGs), and DEGs (Figure 4.7 A). Note that a small number of DETEs, especially DETEs within 2kb flanking regions of genes, might co-localise with multiple DEGs and *vice versa*. Instead of arbitrarily excluding DETEs or DEGs falling into this scenario, the comparison of the expression pattern of co-localized DETEs and DEGs was conducted on each DETE-DEG pair. The expression pattern of the 45 DETEs co-localized with 40 DEGs in mock treatments was then compared with that of paired DEGs, resulting in 45 pairs of DETE-DEG comparisons that were summarised in Figure 4.7 B, where 42 (93.33%) pairs of co-localized DETE-DEG showed concordant clustering between DETEs and corresponding DEGs. The same approach was applied on DETEs and DEGs of yeast and *Botrytis* treatments. In yeast treatment, 20, 68, and 124 genic DETEs of the total 291 DETEs were respectively co-localized with inactive genes, non-DEGs, and DEGs (Figure 4.7 C). There were 106 DEGs of yeast treatment associated with the 124 DETEs co-localized with DEGs, resulting in 126 co-localized DETE-DEG pairs, of which 113 (89.68%) pairs shared the same expression pattern (Figure 4.7 D). With *Botrytis* treatment, 10, 17, and 67 DETEs of the total 116 DETEs were respectively co-localized with inactive genes, non-DEGs and DEGs (Figure 4.7 E). There were 68 pairs of co-localized DETE-DEG, which included 67 DETEs and 54 DEGs. Fifty-five (80.88%) of the 68 co-localized DETE-DEG pairs were found responding to the *Botrytis* treatment with the same expression pattern (Figure 4.7 E, F). These findings show that the dynamic expression pattern of DETEs co-localized with DEGs tended to resemble that of the paired DEGs.

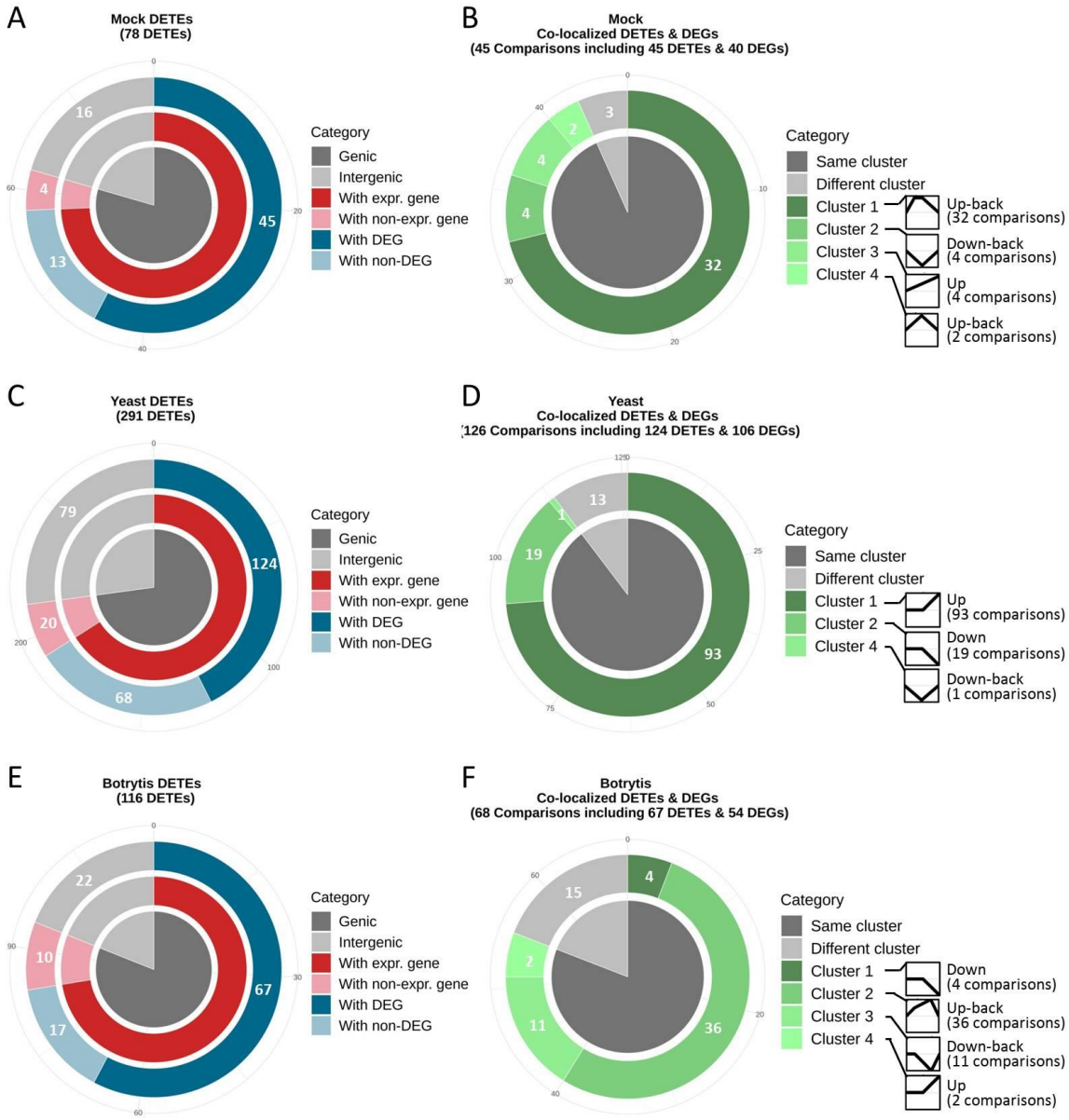



**Figure 4.7 DETE categorization by the activity of co-localized genes and test of expression patterns of co-localized DETE-DEG paired.**

DETEs in mock (A), yeast (C), and *Botrytis* (E) treatments were categorized hierarchically by the presence or absence of co-localized genes (genic or intergenic), transcriptional activity of co-localized genes, and the differential test of expressed genes. The co-localized DETEs and DEGs were further tested for shared expression pattern using hierarchical clustering, which was summarized in pie graphs shown in (B), (D), and (E), respectively, for mock, yeast and *Botrytis* treatments. The expression patterns of the synchronized DETE-DEG pairs were as illustrated in the line graphs.

## 4.5 Discussion

### 4.5.1 A considerable proportion of genes co-localized with TEs in grapevines

Transposable elements have been found to be associated with various proportions of genes in different eukaryotic genomes. In maize, Schnable et al. (2009) found that 66% of genes are sitting within 1 kb of an annotated TE. *Oryza sativa,* with at least 35% of TE content (International Rice Genome Sequencing Project and Sasaki, 2005), was reported to have > 10% of genes overlapping with TEs (Sakai et al., 2007). In humans, it has been reported that almost 25% of promoter regions contain TE-related sequences (Jordan et al., 2003), and approximately 4% of genes have TEs within protein-coding regions (Nekrutenko and Li, 2001). In our study, chapter 3 and this chapter thoroughly interrogated the distribution of TEs in both gene unit and 2kb-flanking region of genes in grapevine and reported this distribution from both perspectives of TEs and genes. As presented in chapter 2, the annotated TE sequences occupy 33% of *V. vinifera* reference genome (Table 2.1). By contrast, as shown in this chapter, the annotated gene exons and introns contribute 10.3% and 23.9% of the genome sequence, respectively. The annotated territories of these TE loci and gene units (including exons and introns) substantially overlap with each other in the reference genome since 33.3% (10,598 genes) of the 31,845 grapevine genes contain 21.5% (48,084 TE loci) of the total 223,411 TE loci within gene unit (Table 3.3 and Table 4.1). Furthermore, another 49.5% (15,754 genes) of grapevine genes are within 2 kb distance of 21.6% (48,351 TE loci) of the total annotated TE loci (Table 3.3 and Table 4.1). Therefore, over 80% of genes in the grapevine either contain TEs or are within 2Kb of a TE insertion. Such a high proportion of grapevine genes associated with TEs suggest that, despite many TEs being located in large sections of gene-poor regions, noticeable proportion of TE loci are closely associated with significant percentage of genes, which in turn implies the substantial participation of TEs in gene regulation.

### 4.5.2 Intragenic TE insertion and TE integrity negatively associate with gene expression level

Although it has been reported that the presence of TEs within genes or close to genes has a negative correlation with the expression level of the co-localized genes (section 4.2; Hollister and Gaut, 2009; Le et al., 2015), possibly due to the epigenetic silencing marks on these TEs, the expression level of genes in association with TE characteristics (e.g. TE expression activity, integrity and location), which might correlate with the level of epigenetic silencing strength posed on TEs, was not thoroughly reported yet. In order to gain a picture of the influence of the presence of TEs near gene coding sequences and the impact of the characteristics of these TEs on co-localized genes with respect to gene expression, the FPKM level of expressed genes without TEs was compared with those co-localized with TEs, with TE transcriptional activity, integrity and location taken into consideration.

In the situation that all of the co-localized TE loci were not expressed, genes harbouring unexpressed TEs in the gene unit showed significantly lower FPKM level than those without TEs (Figure 4.2 A, B), whereas unexpressed TE insertions in the flanking regions had no impact on the expression level of co-localized genes (Figure 4.2 C, D), suggesting that unexpressed intragenic TEs are negatively related with gene expression level. This observation is concordant with that of Le et al. (2015). They found that the expression level of *Arabidopsis* genes without intragenic TEs is significantly higher than that of those having intragenic TEs, and this difference is associated with the degree of CHG methylation on TE feature, which is predominantly maintained by the chromomethylase CMT3 during DNA replication or through DDM1-mediated heterochromatin silencing (Table 1.1).

In the situation that genes within 2kb distance of full-length expression candidates (Figure 4.2 G), the expression level of these genes was significantly lower than that of genes without TEs and that of genes within 2kb distance of full-length unexpressed TEs (Figure 4.2 C). If the TE-expression candidates in the 2kb-flanking regions of genes were all fragmented, then no difference in the expression level of genes was observed (Figure 4.2 H). These findings might be able to be explained by the "expression-dependent RdDM" silencing pathway proposed by Panda et al. (2016), in which transcripts of full-length autonomous TEs are preferentially processed into 21 to 22nt secondary siRNAs that facilitate the transition from PTGS (post-transcriptional gene silencing) to the downstream of canonical RdDM pathway to suppress TE transcription (see section 1.3.2). This pathway would predominantly result in CHH methylation mediated by the DNA methyltransferase DRM2 (Panda et al., 2016). Once the DNA methylation is established, suppressive heterochromatic marks can be deposited on histones by H3K9 methyltransferase and subsequently promote the formation of heterochromatin (Cuerda-Gil and Slotkin, 2016), therefore likely prohibit genes near these heterochromatic regions to be highly expressed.

Likewise, genes containing full-length expression candidates in the gene unit were less transcribed than genes without TEs (Figure 4.2 E), suggesting that these genes (containing full-length expression candidates) might be affected by the "expression-dependent RdDM" pathway that preferentially targets expressed full-length TEs. Intriguingly, we have previously shown that the expression level of genes containing unexpressed full-length TEs in the gene unit was significantly lower than that of genes without TEs (Figure 4.2 A), but the presence of full-length expression candidates in the gene unit (Figure 4.2 E) did not associate with a gene expression level lower than the situation when all full-length intragenic TEs were unexpressed (Figure 4.2 A; p-value = 0.5627). This lack of synergistic negative effect on these two groups of genes, which were different in the transcriptional activity of corresponding intragenic full-length TEs, suggests that these genes might be preferentially affected by different epigenetic silencing pathways respective to the expression activity of co-localized TEs.

This assumption suggests that genes containing unexpressed full-length intragenic TE loci might be predominantly affected by an epigenetic silencing pathway independent of the expression activity of intragenic TEs (e.g. canonical RdDM, CG/CHG methylation during DNA replication, and DDM1-mediated chromatin silencing; also see section 1.3), and this pathway possibly prefers targeting TEs within genes over TEs in the flanking regions. On the other side of this assumption, genes containing full-length expression candidates might tend to be influenced by the "expression-dependent RdDM" pathway that preferentially targets expressed full-length TE loci.

The aforementioned assumption, in which the transcriptional activity of co-localized TEs predispose host genes to different epigenetic silencing pathways, might also explain the significant difference between the two groups of genes wherein the intragenic TEs were all fragmented but different in transcriptional activity (Figure 4.2 B, F; p-value < 0.0005). Genes containing unexpressed fragmented TE loci within the gene unit might be predominantly affected by the epigenetic silencing pathway targeting silenced intragenic TEs, and therefore the expression level of these genes was significantly lower than that of genes without TEs (Figure 4.2 B). By contrast, genes with intragenic fragmented expression candidates might be predominantly associated with the "expression-dependent RdDM" pathway that prefers targeting full-length expression candidates over the fragmented expressed ones, and therefore the expression level of these genes was comparable with that of genes without TEs (Figure 4.2 F), and significantly higher than genes containing non-expressed fragmented TEs (Figure 4.2 B). Overall, our findings are concordant with the published *Arabidopsis* results (Hollister and Gaut, 2009; Le et al., 2015), which suggest that TEs are either unlikely to be accumulated in highly expressed genes or the presence of TE insertions within or close to genes may dampen gene expression due to the epigenetic side effect from co-localized TEs (Hollister and Gaut, 2009). Based on these observations, Hollister and Gaut (2009) have proposed a theory that there is an evolutionary "trade-off" regarding TE silencing and the impact on the expression of neighbouring genes, in which epigenetic suppression on TEs may increase the detrimental impact of intragenic TEs on host genes. In extreme cases, if the deleterious side effect from the strong epigenetic silencing of an intragenic TE insertion severely diminishes expression of an essential host gene, this TE insertion might be subjected to purifying selection, which, in this case, acts against the TE insertion that is epigenetically detrimental to the fitness of individuals by purging this TE insertion from population and results in highly conserved TE-insertional pattern in this genic region pivotal to fitness (Chuong et al., 2016). On the contrary, if the TEs within or proximal to genes do not genetically or epigenetically disrupt the expression of the host genes, these TE insertions are likely to be retained in the population (Hollister and Gaut, 2009). In this case, as a "trade-off", the expression activity of these host genes is secured (possibly at a low transcription level) due to relatively relaxed epigenetic suppression and chromatin status of these sites compared with completely silenced heterochromatic

region, while the co-localized TEs, in turn, have access to transcription machinery due to the permissive transcriptional status granted to the host genes. . This “trade-off” theory implies that it is likely the function or necessity of the genes that drive the association between the expression activity of genes and the epigenetic status of co-localized TEs. This evolutionary trajectory may eventually result in an epigenetic landscape where the activity of intragenic TEs is fine-tuned to maintain minimum impact on host genes, and therefore explains the skewed distribution of TE expression candidates toward intron of expressed genes (chapter 3).

### 4.5.3 Expression patterns of DETEs mostly concordant with that of co-localized DEGs

Following the “trade-off” theory mentioned in the previous section, although the presence of TE insertions within or proximal to genes might prohibit these host genes from being highly expressed possibly due to the epigenetic side effect from the co-localized TEs, the epigenetic suppression level posed on these TEs might have been fine-tuned to allow the minimum required transcriptional activity of the host genes because of the necessity of these gene transcripts. This relaxed epigenetic silencing status granted to host genes might, in turn, benefit the co-localized TEs in terms of the chance to be transcribed. Under this assumption, the dynamic transcriptional pattern of TEs co-localized with expressed genes is possibly concordant with that of host genes, providing that these TEs and the co-localized genes share similar chromatin status. In order to define temporal patterns of TE and gene expression with statistical power, we were only able to interrogate TEs identified as “trackable expression candidates” that were mapped with unique-mapping reads (see sections 4.2 and 4.3.2). Although this approach largely reduced the number of TEs that are available for the investigation, it allowed the hierarchical classification of expression pattern that facilitates the comparison between TEs and genes.

The hierarchical clustering on paired DETEs and DEGs reveals that most co-localized DETE and DEGs were grouped into the same clusters based on their expression patterns (Figure 4.7 B, D, F). As there existed various expression patterns in different treatments (i.e. “up-back” pattern in mock, “up” pattern in yeast, and “up-back” in *Botrytis* treatments), the concordance of the dynamic activity between DETEs and DEGs is not limited to specific expression pattern. The concordant transcriptional dynamics between DETEs and co-localized DEGs were also observed in TEs near up-regulated genes in response to abiotic stress in maize (Makarevitch et al., 2015). In *Arabidopsis*, many DETEs responding to salicylic acid shared similar transcriptional alterations with co-localized DEGs (Dowen et al., 2012). They also found that the up-regulation of TEs is associated with demethylation of the underlying DNA. Using time-series experiments, *Arabidopsis* experiencing phosphate starvation revealed gradually increased CHH methylation in regions overlapping with genes that were

dynamically activated across time (Secco et al., 2015). Interestingly, the hypermethylation changes in these differentially methylated regions (DMRs) occurred after the initiation of gene activation in response to this stress. Moreover, almost all of these hypermethylated DMRs overlapped with TEs. It should be noted that PTGS can crosstalk with canonical RdDM and that this can eventually lead to DNA methylation on TE content having transcriptionally active history (Cuerda-Gil and Slotkin, 2016), and a rebound of CHH methylation has been observed in response to activation of *EVD* retrotransposon in *Arabidopsis* (Marí-Ordóñez et al., 2013). Therefore, in Secco et al. (2015), it is possible that, in the same DMRs, the CHH hypermethylation initiated after gene activation implicates the attempt to re-silence TEs co-activated with the co-localized up-regulated genes. In cases showing the consistent expression pattern of co-localized TEs and genes, it is not certain that whether the stress-responsiveness of TEs leads to similar transcriptional alteration in co-localized genes, or TEs take advantage of the dynamic chromatin state for gene expression. One scenario may apply to TEs at the promoter regions while the other could be adopted by intronic TEs. However the commonality between the two is the permissive chromatin structure that allows transcription taking place. Therefore, as a possible explanation for the similar transcriptional behavior of co-localized DETEs and DEGs, TEs and co-localized genes sitting within genomic regions possessing similar chromatin states during stress response may exhibit similar dynamic transcription.

## 4.6 Conclusions

Although TEs are littered across the grape genome, there are a high proportion of grapevine genes that contain TEs in their exons, introns, and 2kb flanking regions. This implies a close evolutionary and regulatory connection between TEs and genes. Our findings are concordant with the observations reported by Hollister and Gaut (2009) and Le et al. (2015), where the presence of TE insertions within or neighbouring a gene appear to dampen the transcriptional activity of that gene. Moreover, our analysis further distinguished the possible impact of the transcriptional activity, integrity and location of TEs on the transcriptional level of host genes. In General, full-length TEs, gene body insertions, and transcriptionally active TEs are less likely to associate with highly expressed genes. In fact, the results from this analysis imply that the transcriptional activity of co-localized TEs might predispose host genes to different epigenetic silencing pathways that preferentially target TEs of other different characteristics; while genes co-localized with expression candidates are likely to be predisposed to the "expression-dependent RdDM" pathway if the co-localized active TEs are full-length, genes co-localized with unexpressed TEs might be predominantly affected by an epigenetic silencing pathway that favours targeting TEs in gene unit over TEs within 2kb distance of genes.

Although gene activity was found to be negatively associated with the presence of co-localized TEs likely due to the epigenetic side effect from these TEs, the co-localized TEs might take advantage of the permissive transcriptional status of the genic regions for their expression activity, and therefore display expression dynamics resemble to host gene's activity. The comparison of the expression pattern between co-localized DETEs and DEGs shows that there is a tendency toward concordant transcriptional dynamics, suggesting a similar chromatin state around individual pair of DETEs and DEGs.

This research, for the first time, addresses the two phenomena (i.e. the negative correlation between gene expression and presence of TE insertions versus concordant expression pattern of DEGs and co-localized DETEs) that have been previously reported but appear contradictory. We, therefore, hypothesise that these two phenomena might be linked with the "trade-off" theory proposed by Hollister and Gaut (2009).

The data of chapter 3 and chapter 4 reveal the closely connected transcriptional relationship between TEs and co-localized genes. It is possible that the location distribution of expressed TE loci relative to genes varies between genotypes and species. To examine this assumption, in the next chapter, our analysis workflow was applied on published short-read RNAseq data of *A. thaliana*,

including wild-type and mutant deficient in epigenetic silencing components, as well as genetically transformed *Drosophila melanogaster* that has shown a storm of TE transcription.

# Chapter 5

# Application of the new analysis pipeline in *Arabidopsis* and *Drosophila* RNAseq data

## 5.1 Overview

Although TE sequences account for 223,411 annotated loci which occupy 30-40% of the *Vitis vinifera* genome, only 1.7 % of these loci were found to be potentially active in grapevine embryogenic callus. Less than 2.5 % of the total annotated loci appeared to be transcriptionally active when the callus was exposed to wound-like and biotic stress treatments. These loci were discovered via the analysis pipeline described in chapter 2 that collected potentially expressed TE loci or expression candidates. The application of this pipeline further revealed that the expression candidates were predominantly positioned within the intron of expressed genes (chapter 3).

To test whether this analysis approach has utility for analysing RNAseq data of other species and examine whether the location bias we observed is a common phenomenon, the published RNAseq datasets generated from *Arabidopsis thaliana* and *Drosophila melanogaster* were analysed using this workflow. As expected, while 4.5% and 4.3% of the total annotated *A. thaliana* TEs were identified as expression candidates in wild-type and the *ibm2* mutant, respectively, the number of expression candidates increased over threefold in *Arabidopsis* mutants that were deficient in the chromatin re-modeller gene *DDM1*. Location and distribution of wild-type and *ibm2* TE expression candidates confirmed the location bias towards expressed genes observed in grapevine, whereas the proportion of intergenic expression candidates was dramatically elevated in *ddm1*, concordant with DDM1's suppressive property on heterochromatic TEs. As a positive control, TE loci derived from *ATCOPIA93* (*EVD*), which has been proved to transpose in *ddm1*, were identified as origins of autonomous TE transcripts. In comparison, about 50% of *D. melanogaster* annotated TEs were potentially expressed in the *Drosophila* ALS model, where a TE storm has been proposed to be stimulated by ectopic expression of human TDP-43 in the glial and neuronal cells. Using our approach, the TE storm was de-convoluted to identify active elements at the individual TE level. While the distribution of expression candidates was in the main maintained as same as the distribution of all annotated loci, the proportion of intronic expression candidates within active genes was significantly increased in the glial model of ALS compared with healthy flies and the neuronal model of ALS. The identification of these individual TE loci and the co-localized genes by our analysis approach might facilitate the understanding of ALS pathogenesis.

## 5.2 Introduction

Using the analysis pipeline established in chapter 2, a small subset of annotated TE loci was identified as expression candidates that potentially expressed in grapevine embryogenic callus with stress treatment. Characterization of these TE loci revealed a strong location bias towards introns of expressed genes, leading to the assumption that TEs positioning in the intron of active or inducible genes are more likely to be transcriptionally active (see chapter 3). To understand whether these findings in grapevines is an exceptional case or is conserved with other species, published RNAseq data derived from epigenetically compromised *Arabidopsis thaliana* mutants and neuronal degenerated *Drosophila melanogaster* models exhibiting TE storms were analysed as per grapevine.

Firstly, the function of the analysis pipeline and the characteristics of expression candidates described in chapter 2 and chapter 3 was tested in RNAseq data of *Arabidopsis* wild-type and *ibm2*, which is a mutant that is incompetent in preventing the use of the intrinsic polyadenylation sites derived from intronic TEs but does not exhibit large scale of TE activation (also see section 5.2.1 for detail; Deremetz et al., 2019; Ito et al., 2016; Saze et al., 2013; Wang et al., 2013). The *Arabidopsis ddm1*, a mutant that is compromised in epigenetic silencing of TEs and thus suffered from extensive TE mobilization (Oberlin et al., 2017), was used as a positive control of TE activation (see section 5.2.1 for detail). Analysing the RNAseq data of *ddm1* can demonstrate whether our analysis workflow can effectively identify expression candidates whose characteristics (e.g. integrity and location) are concordant with transposons shown to mobilise due to the loss of *DDM1* epigenetic silencing.

Secondly, our analysis pipeline was applied on the RNAseq data of neuronal degenerated *D. melanogaster* models (Krug et al., 2017) exhibiting TE storms (see section 5.2.2 for detail). Overexpression of Gypsy elements, determined at the family level, was found to be a potential cause of the pathological symptoms of these *Drosophila* models (Krug et al., 2017). However, the individual active loci of these TE families remain unknown. We, therefore, tested whether the use of our analysis pipeline for this dataset could demonstrate increased granularity of the current analysis to uncover the individual transcriptionally active TE loci in this system.

### 5.2.1 The function of DDM1 and IBM2 in *A. thaliana*

TE transposition has been detected in *Arabidopsis* depleted in the chromatin re-modeller DDM1 (Decrease in DNA Methylation 1; Lee et al., 2020; Tsukahara et al., 2009). DNA methylation on TE sequences is a key epigenetic mechanism to suppress TE activity. Generally speaking, in plants, there are three types of methylated sequence contexts; CG, CHG, and the asymmetric CHH, where H denotes any nucleotide except G. In plants, Methyltransferase 1 (MET1) typically executes the placement of methyl groups on the cytosine of the CG couplet in TE and gene body contexts (Law

and Jacobsen, 2010). CHG and CHH methylation are more prevalent in TE DNA or repetitive sequences than in other parts of the genome. Chromomethylase 3 (CMT3) catalyses the deposition of methyl groups in the CHG context while CMT2 and Domains Rearranged Methyltrasferase 2 (DRM2) mediate CHH methylation (Springer and Schmitz, 2017). Although the function of DRM2 and CMT2 appear redundant, the former is predominantly responsible for CHH methylation on short TEs (e.g. < 1kb) and TE boundaries through the RNA Dependent DNA Methylation (RdDM) pathway, while the latter specializes in CHH methylation on long TEs, targeting, in particular, the internal sequences of the TE, through a DDM1-dependent pathway that is distinct from RdDM (Springer and Schmitz, 2017; Zemach et al., 2013). DDM1 is homologous to the yeast Snf2 protein family that transforms the energy derived from ATP hydrolysis into the physical alteration of nucleosome composition and chromatin structure (Ryan and Owen-Hughes, 2011). In a higher order of nucleosome assembly, histone H1 bridges linker DNAs that intersperse nucleosomes and therefore stacks nucleosomes into a more compact structure that characterises heterochromatin (Bednar et al., 2017; Misteli et al., 2000). It is believed that H1 prevents access of the RdDM machinery to tightly packed heterochromatin, whereas the presence of DDM1 enables the accessibility of the H1-bound heterochromatin to DNA methyltransferases, particularly CMT2 (Zemach et al., 2013). An increase in TE insertions derived from several LTR-TEs, including *ATCOPIA93* (*EVD*), was observed in the *Arabidopsis* inbreed lines deficient in *DDM1* (Tsukahara et al., 2009). This is concordant with *EVD*'s transcriptional activation (Oberlin et al., 2017) and elevated accumulation of VLP-enclosed cDNA one step prior to the re-insertion into the host genome (Lee et al., 2020). It is, therefore, expected that analysis of unmasked RNAseq sequence data from the *ddm1* mutant, using the analysing approaches described in chapter 2 and chapter 3, would reveal a wide range of TE activation indicated by the increased number and elevated intergenic proportion of expression candidates.

The epigenetic regulation of intronic TEs seems more complicated than the intergenic TEs. As mentioned in chapter 1, unmasked intronic TEs can interfere with gene transcript splicing, whereas over-expansion of the silencing marks (e.g. methylated CHG/CHH and H3K9me2) from intronic TEs into a gene's coding regions has the potential to suppress gene transcription (Ong-Abdullah et al., 2015; Sigman and Slotkin, 2016; Tsuchiya and Eulgem, 2013). Key factors to prevent silencing marks extending from intragenic TEs include the histone demethylase IBM1 (INCREASE in BONSAI METHYLATION 1) that removes di-methyl groups from H3K9 and the bi-functional DNA glycosylase/lyase ROS1 (REPRESSOR OF SILENCING 1) that harnesses its glycosylase function to remove methylcytosine from gene coding region (Saze et al., 2008; Zhu et al., 2007). In addition to the enzymes that remove epigenetic silencing marks from the coding region, the RNA binding protein IBM2 was found to be necessary for genes having intragenic TEs to avoid the use of the intrinsic polyadenylation sites derived from the intronic TEs (Deremetz et al., 2019; Ito et al., 2016; Saze et al.,

2013; Wang et al., 2013). Nonetheless, heterochromatic DNA methylation was still maintained in *ibm2*, indicating that *IBM2* mainly regulates correct splicing instead of directly driving transcriptional silencing of intronic TEs (Le et al., 2015; Saze, 2018). Therefore it is expected that the number and location distribution of expression candidates in *ibm2* would be as similar as that found in wild-type.

### 5.2.2 Epigenetic silencing of TEs in *D. melanogaster*

In comparison to plants, DNA methylation is rarely found in the *D. melanogaster* genome. Instead, the animal-specific Piwi-interacting RNAs (piRNAs) and endogenous siRNAs are crucial for TE suppression in both post-transcriptional gene silencing (PTGS) and transcriptional gene silencing (TGS), which mostly refers to the accumulation of H3K9me2 at chromatin regions containing TEs (Czech and Hannon, 2016; Mérel et al., 2020). In the ping-pong silencing loop, piRNA precursors transcribed from genomic piRNA clusters by Pol II are processed by the Argonaute protein Ago3 into 23-30 nt antisense piRNAs. The Aubergine (Aub) protein then carries the antisense piRNAs to the complementary TE transcripts, leading to the digestion of TE mRNAs and generation of sense piRNAs that navigate Ago3 to the piRNA precursors, and thus forming a forward-feeding loop (Czech and Hannon, 2016; Huang et al., 2017). Alternatively, the piRNA precursors can be recognized by the Piwi protein, which mediates the synthesis of phased piRNAs (Mérel et al., 2020). In the nucleus, piRNAs produced through both the ping-pong cycle and phased piRNA pathway can guide Piwi to the genomic TE insertions and facilitate H3K9 methylation. It is believed that the ping-pong cycle only occurs in *Drosophila* germ cells, while phased piRNA biogenesis takes place in germ cells and the ovarian somatic follicle cells (Czech and Hannon, 2016; Huang et al., 2017; Mérel et al., 2020). On the contrary, the siRNA-mediated PTGS is not limited to germ cells and the ovarian somatic. In *D. melanogaster*, dsRNAs can be processed by the Dicer protein Dcr-2 into siRNAs, which would then navigate Ago2 to TE transcripts with sequence complementarity, resulting in cleavage of TE mRNAs (Mérel et al., 2020).

Stress-induced transposition bursts have been reported in *Drosophila,* with specific TE family activation dependant on the genetic background (Guerreiro, 2012; Mérel et al., 2020). Interestingly, Krug et al. (2017) observed a Gypsy TE storm in the amyotrophic lateral sclerosis (ALS) *Drosophila* model with ectopic expression of human TAR DNA-binding protein 43 (hTDP-43). Cytoplasmic aggregation of this protein is thought to be the major pathological signature and cause of a large proportion of human ALS (Krug et al., 2017; Wang et al., 2008). In this *Drosophila* model, large scale TE transcriptional activation was detected due to the hTDP-43 over-expression in the neuronal and glial cells, where the activity of Gypsy was of particular concern for that de-repression and mobilization of this TE superfamily have been observed in advanced-aged brain tissue of *Drosophila* (Krug et al., 2017). Introduction of siRNAs against Gypsy transcripts and pharmacological inhibition of

retroelement's reverse transcriptase showed substantial improvement of the lifespan of individual flies in this model. In addition to Gypsy elements, several other TE families were also transcriptionally activated in these flies. As these studies only analysed TE activity at the family level, we hypothesised that our analysis approach described in chapter 2 and chapter 3 would contribute to the identification of individual active TE loci.

## 5.3 Methods

### 5.3.1 Acquisition and analysis of the *A. thaliana* and *D. melanogaster* RNAseq data

RNAseq data of *A. thaliana* wild-type and *ibm2* was obtained from Le et al. (2015; accession codes DRA002305 and DRA002306 in DDBJ Sequence Read Archive at https://www.ddbj.nig.ac.jp/dra/), while the *ddm1* RNAseq data was collected from Oberlin et al. (2017; accession code GSE93584 in NCBI Gene Expression Omnibus at http://www.ncbi.nlm.nih.gov/geo/). RNAseq data of *D. melanogaster* TDP-43 ALS model was acquired from Krug et al. (2017; accession code GSE85398 in NCBI Gene Expression Omnibus at http://www.ncbi.nlm.nih.gov/geo/). The *A. thaliana* TAIR10 reference genome, tRNA/rRNA sequences, and gene annotation file were downloaded from Ensembl Plants (https://plants.ensembl.org), whereas the corresponding TE annotation file was generated by Jin et al. (2015) and is available from Prof. Molly Hammell's lab web page (http://hammelllab.labsites.cshl.edu/). The *D. melanogaster* dm3 reference genome was collected from the FTP site of UCSC Genome Browser (https://genome.ucsc.edu/cgi-bin/hgGateway). The dm3 rRNA sequences, gene annotation and TE GTF file were downloaded from Prof. Molly Hammell's lab web page, while tRNA sequences were obtained from FlyBase (https://flybase.org/). Data pre-processing, alignment, as well as identification and characterization of expression candidates were performed as described in chapter 2 and chapter 3.

## 5.4 Results

### 5.4.1 Collection of *A. thaliana* expression candidates using the pipeline

Transcriptome data from Le *et al.* (2015) and Oberlin *et al.* (2017) were analyzed as described in chapter 5.3 Methods. In concordance with Le *et al.* (2015) and Oberlin *et al.* (2017), the mapping statistics (Table 5.1) showed roughly 80 to 95% of sequenced reads were mapped to the unmasked TAIR10 *A. thaliana* reference genome. Following the analysis pipeline used in previous chapters, 1,410 (4.52%) and 1,342 (4.30%) expression candidates were identified from 31,189 annotated TEs in wild-type *Col* and *ibm2,* respectively (Figure 5.1 A, B). In contrast, over 4,000 TEs were found potentially expressed in the *ddm1* mutant (Figure 5.1 C). Furthermore, while untrackable candidates contributed to less than 6 % of the candidate pool in wild-type and *ibm2* (Figure 5.1 D, E), the proportion of untrackable candidates reached 12.9% in *ddm1* (Figure 5.1 F). Analyzing at the family level, 228 of the 320 TE families in the TAIR10 genome obtained expression candidates in the wild-type plant (Figure 5.1 G). As *ibm2* obtained a similar number of active TE families (224 TE families) identified by the new pipeline, *ddm1* revealed 292 active TE families collected through the same method (Figure 5.1 H, I). By use of the software TEtranscripts, 159, 146, and 273 TE families were, respectively, found active in wild-type, *ibm2*, and *ddm1*, which were largely overlapping with the corresponding subsets identified by the new pipeline (Figure 5.1 G-I).

**Table 5.1 Mapping statistics for RNA-seq analysis of *Arabidopsis* dataset of Le *et al.* (2015) and Oberlin *et al.* (2017).**

| Sequenced libraries | | Sequenced reads | | Adaptor removal | | Filter tRNA/rRNA | | Mapped reads | |
|---|---|---|---|---|---|---|---|---|---|
| Genotypes | Replicates | | | | | | | | |
| Wild-type | a | 182,422,324 | 100% | 173,591,432 | 95.16% | 172,988,410 | 94.83% | 172,185,257 | 94.39% |
| *Ibm2* | a | 143,454,238 | 100% | 136,154,752 | 94.91% | 135,787,848 | 94.66% | 134,833,956 | 93.99% |
| *ddm1* | a | 148,929,796 | 100% | 148,876,038 | 99.96% | 127,741,608 | 85.77% | 118,686,674 | 79.69% |
| | b | 137,374,464 | 100% | 137,324,766 | 99.96% | 134,179,866 | 97.67% | 125,908,723 | 91.65% |

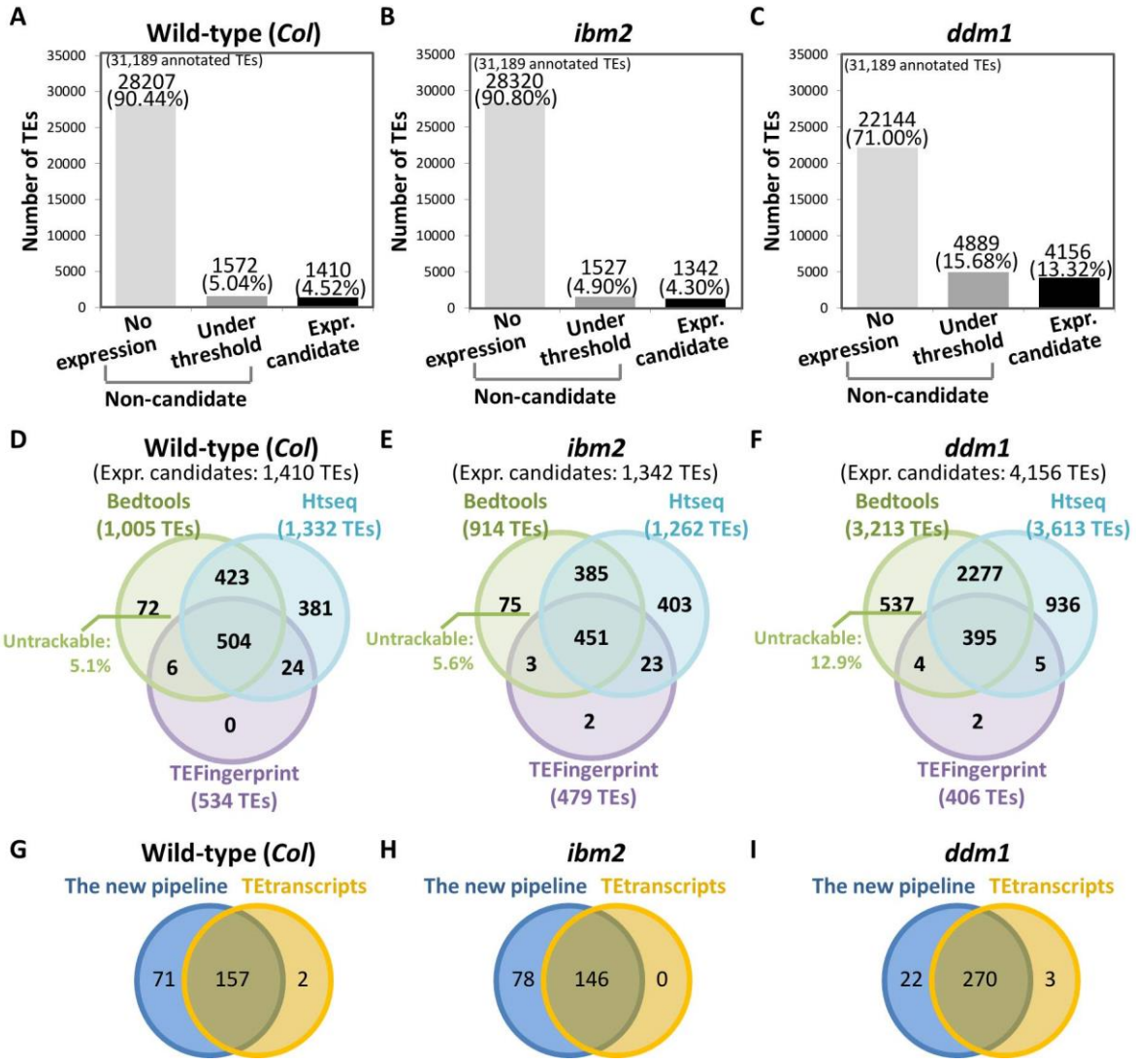



**Figure 5.1 Expression candidates of wild-type (*Col*), *ibm2* (Le et al. 2015) and *ddm1* (Oberlin et al. 2017) *Arabidopsis* identified by the pipeline**

**(A)-(C)** All annotated TEs were categorized by transcriptional activity and illustrated according to treatment as indicated. **(D)-(F)** The expression candidates (expr. candidates) of each treatment was a pool of potentially expressed TEs collected by the three sub-pipelines. **(G)-(I)** The expr. candidates were grouped by family to identify transcriptionally active TE families, which were then further compared with the active families captured by TEtransctipts.

Despite the high consistency between the collections of expressed TE families from our method and TEtranscripts, some TE families were only collected by one or the other. The two families uniquely identified by the TEtranscript approach in the wild-type data contained 32 individual loci, of which 25 loci had zero expression. The expression levels of the remaining seven loci were either lower than our threshold of 10 read counts or fell under 5 average depth of mapped region (Figure 5.2 A). Although the total reached the expression threshold set for the TEtranscripts approach at a family level, none of the individual loci was considered expressed due to low count and read coverage revealed by the new pipeline. For the same reason, the 3 TE families uniquely included by TEtransciprt in *ddm1*, where only 38 of the 175 loci from the three families were mapped, yet they

were all under the expression threshold (Figure 5.2 B). On the other hand, although 71, 78, and 22 TE families were only identified by the new pipeline in the wild-type, *ibm2* and *ddm1*, respectively, the expression candidates were collected by the pipeline due to one of the three situations shown in Figure 5.3:

(1) Situation 1: the TE loci were mapped by reads across the respective junctions between TEs and exons of genes;

(2) Situation 2: the TE loci were mapped by reads that were also mappable to genes because the TE loci overlapped with the exon of these genes;

(3) Situation 3: The read counts and read coverage of the TE loci were merely above the expression threshold in the new pipeline but failed the threshold of TEtranscripts.

In Situation 1 and 2, reads mapping to both TEs and genes were preferentially assigned to genes by TEtranscripts. The majority of these TE loci were uniquely identified in our pipeline due to Situation 2 (Figure 5.3).

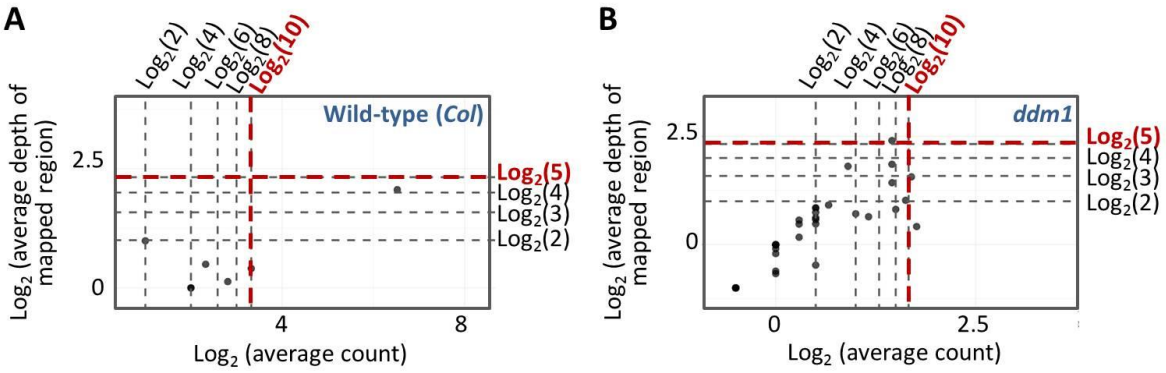



**Figure 5.2 Expression range of individual *Arabidopsis* TE loci from the TE families uniquely included by TEtranscript-based method.**

Each black dot represents a TE locus with the read count value projected on the x-axis and the average depth of mapped region projected on the y axis in **(A)** wild-type and **(B)** *ddm1*. The dual thresholds were indicated by red dash lines.

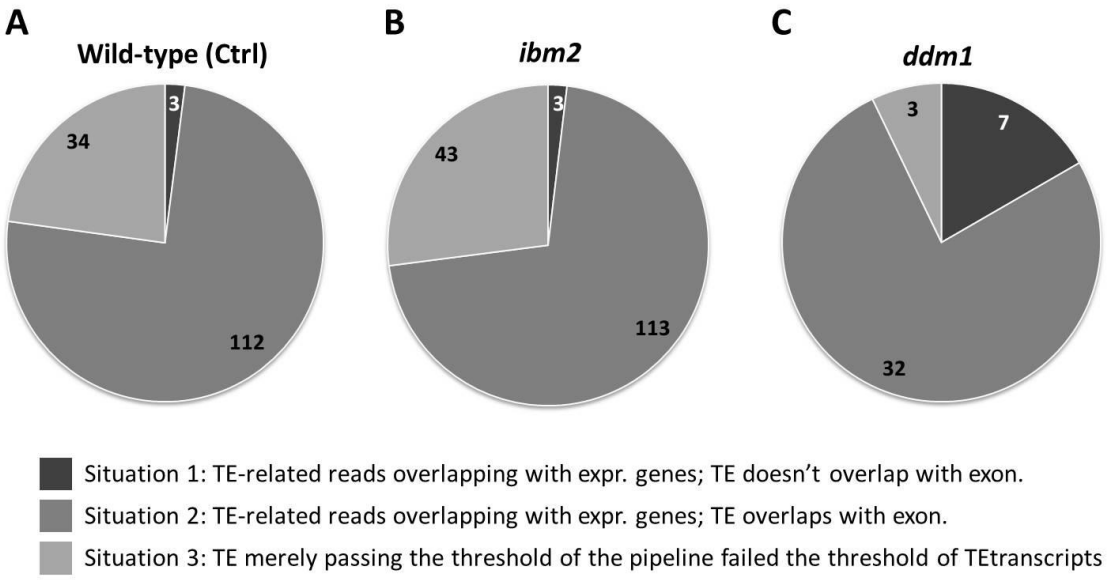



**Figure 5.3 Reasons for the *Arabidopsis* TE families being uniquely included in the new pipeline.**

For each genotype, the expression candidates from the TE families uniquely included by the pipeline were binned into 3 situations indicated. The number of binned TE loci was shown on each slice.

Since the number of expression candidates in *ddm1* was considerably increased, a further comparison among the three sets of expression candidates from the three genotypes was illustrated in Figure 5.4. Most of the expression candidates in wild-type and *ibm2* remained transcriptionally active in *ddm1* (overlapping areas in Figure 5.4 A), whereas an additional 3,091 expression candidates were exclusively detected in the *ddm1* mutant, which is compromised in nucleosome remodelling. Similarly, the comparison of transcriptionally active TEs at a family level showed that the majority of the expressed TE families in wild-type and *ibm2* were also found active in *ddm1*, and a subset with 58 active TE families was uniquely activated in *ddm1* (Figure 5.4 B).

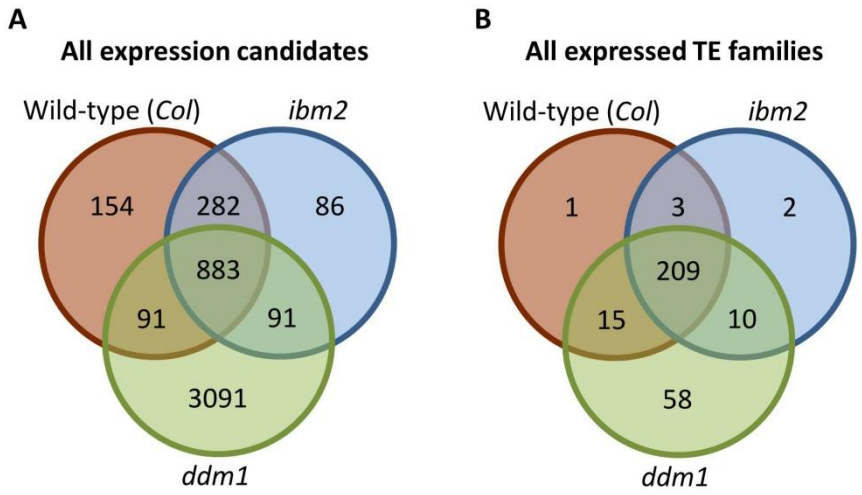



**Figure 5.4 Comparison of *Arabidopsis* expression candidates and active TE families among the three genotypes.**

Comparisons of different sets of **(A)** expression candidates and **(B)** active TE families from the indicated genotypes were illustrated by Venn diagrams, where the overlapping areas denote expression candidates appeared in more than one genotype and the non-overlapping areas denote those exclusively found in specific genotype.

### 5.4.2 The integrity of the *A. thaliana* expression candidates

Among the 31,189 TE loci annotated in the TAIR10 reference genome, 3,842 loci (12.32%) retain > 90% length coverage relative to the canonical sequences (Figure 5.5 A). With the size of 1,300 to 1,400 expression candidates, the proportion of full-length expression candidates in wild-type (11.84%, 167 TEs) and *ibm2* (12.00%, 161 TEs) is similar to that in the all annotated pool (Figure 5.5 B, C). In addition to a lift in the number of expression candidates, *ddm1* mutant showed 706 full-length expression candidates, meaning an increase by 5% compared with wild-type (Figure 5.5 D).

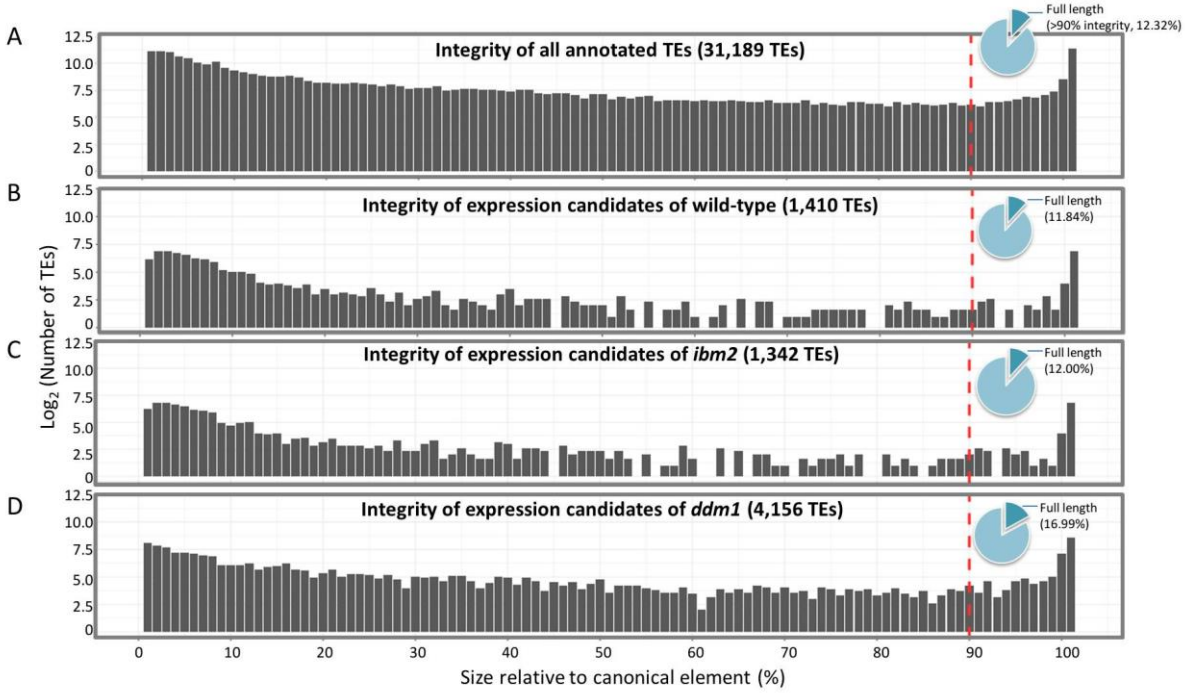



**Figure 5.5 Integrity of annotated *A. thaliana* TEs**

The length of each individual TE locus was compared with the size of the corresponding canonical element. TEs with over 90% length coverage were considered as full length. **(A)** All annotated TEs (31,189 TEs). **(B-C)** Expression candidates of wild-type (B), *ibm2* (C), and *ddm1* (D).

### 5.4.3 Hierarchical classifications of *A. thaliana* expression candidates by location, integrity, and distinctness

As described in chapter 3 and illustrated in Figure 5.6 A, all annotated TEs were firstly grouped by whether they were in the genic or intergenic regions. The former denotes gene unit comprised with exon and intron, and 2kb up and downstream of the gene unit. The rest of the genome was denoted as the intergenic region. This classification showed that 58.37% of *A. thaliana* annotated TEs located in the genic region, while TEs in gene unit and flanking regions, respectively, comprised 16.00% and 42.37% of the total pool (Figure 5.6 B, Table 5.2). Remarkably, the majority of the TEs annotated within the 'gene unit' overlapped with exons (Figure 5.6 B), contributing to 12.42% of all annotated loci.

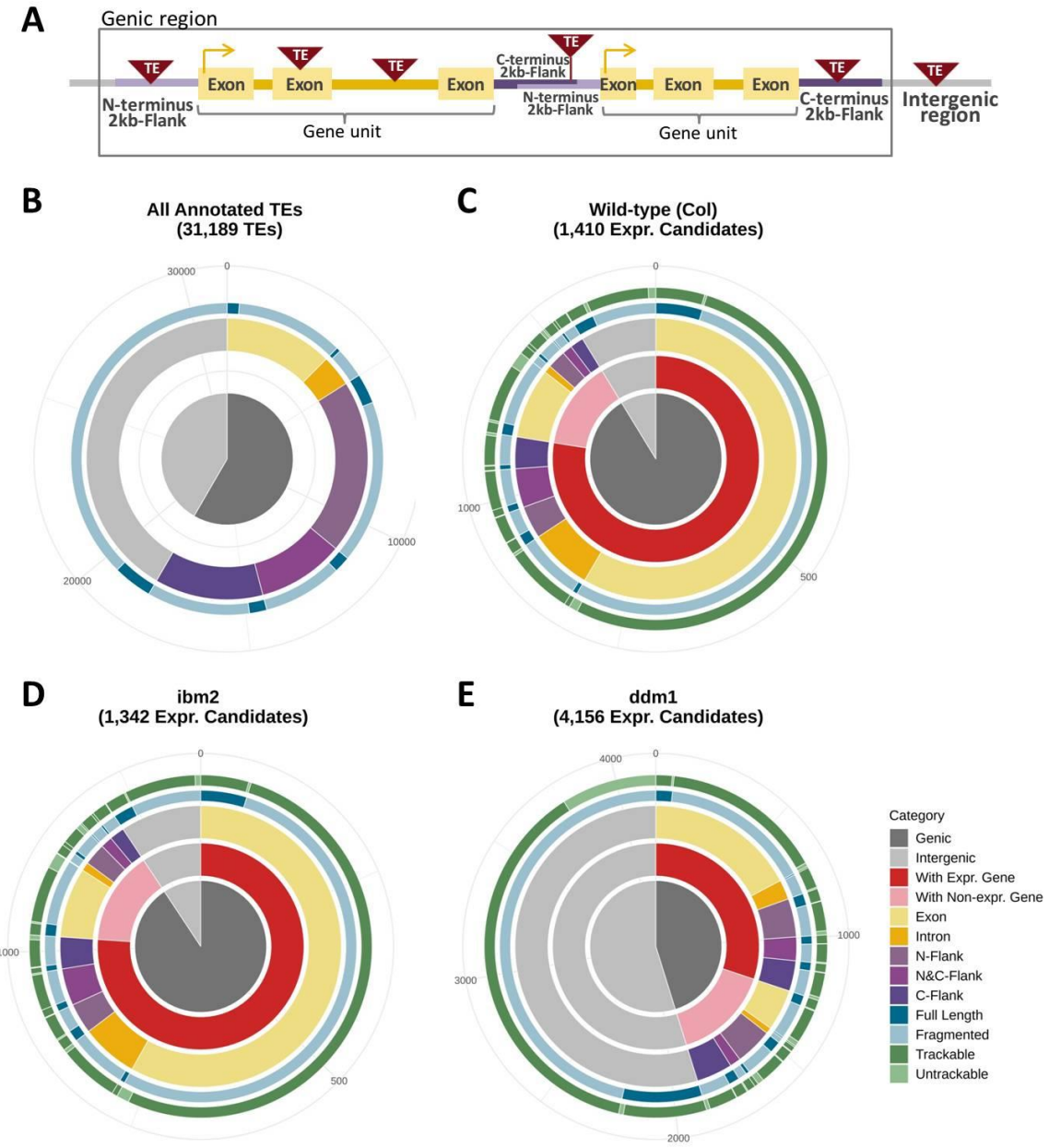



**Figure 5.6 Hierarchical classifications of *A. thaliana* expression candidates by location, integrity, and distinctness.**

**(A)** TEs overlapping with exon, intron, or 2kb upstream (N-terminus) or downstream (C-terminus) of a gene were denoted as genic TEs; otherwise, they were grouped as intergenic TEs. **(B)** All annotated TEs were categorized hierarchically by region (centre), location (internal layer) and integrity (outer-most layer). **(C-E)** Expression candidates of each genotype were categorized in the order of region (centre), the transcriptional activity of co-localized genes (2nd layer), location (3rd layer), integrity (4th layer), and the presence/absence of unique-mapping reads (outer-most layer).

**Table 5.2 Hierarchical categorization of *A. thaliana* annotated TEs by location and integrity.**

Numbers of TEs (#TE) in black colour sum up to 31,189 annotated TEs, while the corresponding percentages (%) in black colour sum up to 100.00%. Subtotals of indicated categories were denoted in grey.

| Region | Location | | All annotated TEs | |
|---|---|---|---|---|
| | | | #TE | % |
| **Genic** | Gene unit | Exon | 3,874 | 12.42% |
| | | Intron | 1,115 | 3.58% |
| | *Subtotal (Gene unit)* | | *4,989* | *16.00%* |
| | Flanks | N-Flank | 6,233 | 19.99% |
| | | N&C-Flank | 3,098 | 9.93% |
| | | C-Flank | 3,883 | 12.45% |
| | *Subtotal (Flanks)* | | *13,214* | *42.37%* |
| *Subtotal (Genic)* | | | *18,203* | *58.37%* |
| **Intergenic** | | | 12,986 | 41.63% |
| *Subtotal (Intergenic)* | | | *12,986* | *41.63%* |
| **Sum** | | | 31,189 | 100.00% |

In wild-type plants, TE loci located in the genic region comprised 91.28% of the 1,410 expression candidates, of which 1,093 loci co-localized with expressed genes (Figure 5.6 C, Table 5.3). In addition, the proportion of gene-unit loci had skewed from 16% of total annotated TE loci (Figure 5.6 B, Table 5.2) to 74.61% of the expression candidates in wild-type plants (Figure 5.6 C, Table 5.3). The majority of these expression candidates in the gene unit were associated with expressed genes (65.67% of the total expression candidates). The *ibm2* expression candidates showed very similar location distribution to the wild-type (Figure 5.6 D). However, while intergenic expression candidates comprised 8.72% of the wild-type candidate pool, the intergenic proportion in *ddm1* was considerably higher at 54.79% of expressed TE loci (Figure 5.6 E, Table 5.3). As opposed to the high proportion of 'gene-unit' expression candidates in wild-type (74.61%), only 24.95% of the expression candidates in *ddm1* were located in the gene unit. These results show that the distribution of expression candidates in *ibm2* mutant is highly similar to that in wild-type, whereas a striking difference was observed between wild-type and *ddm1*.

**Table 5.3 Hierarchical categorization of *A. thaliana* expression candidates by location and integrity.**

Numbers of TEs (#TE) in black colour sum up to 31,189 annotated TEs, while the corresponding percentages (%) in black colour sum up to 100.00%. Subtotals of indicated categories were denoted in grey.

| Region | Gene activity | Location | | Wild-type | | *ibm2* | | *ddm1* | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | #TE | % | #TE | %. | #TE | % |
| **Genic** | With expr. Gene | Gene unit | Exon | 824 | 58.44% | 780 | 58.12% | 717 | 17.25% |
| | | | Intron | 102 | 7.23% | 88 | 6.56% | 94 | 2.26% |
| | | *Subtotal (Gene unit)* | | *926* | *65.67%* | *868* | *64.68%* | *811* | *19.51%* |
| | | Flanks | N-Flank | 53 | 3.76% | 47 | 3.50% | 183 | 4.40% |
| | | | N&C-Flank | 63 | 4.47% | 57 | 4.25% | 113 | 2.72% |
| | | | C-Flank | 51 | 3.62% | 49 | 3.65% | 145 | 3.49% |
| | | *Subtotal (Flanks)* | | *167* | *11.85%* | *153* | *11.40%* | *441* | *10.61%* |
| | *Subtotal (With expr. gene)* | | | *1,093* | *77.52%* | *1,021* | *76.08%* | *1252* | *30.13%* |
| | With non-expr. Gene | Gene unit | Exon | 114 | 8.09% | 108 | 8.05% | 197 | 4.74% |
| | | | Intron | 12 | 0.85% | 13 | 0.97% | 29 | 0.70% |
| | | *Subtotal (Gene unit)* | | *126* | *8.94%* | *121* | *9.02%* | *226* | *5.44%* |
| | | Flanks | N-Flank | 30 | 2.13% | 34 | 2.53% | 177 | 4.26% |
| | | | N&C-Flank | 16 | 1.13% | 18 | 1.34% | 52 | 1.25% |
| | | | C-Flank | 22 | 1.56% | 23 | 1.71% | 172 | 4.14% |
| | | *Subtotal (Flanks)* | | *68* | *4.82%* | *75* | *5.59%* | *401* | *9.65%* |
| | *Subtotal (With non-expr. gene)* | | | *194* | *13.76%* | *196* | *14.61%* | *627* | *15.09%* |
| *Subtotal (Genic)* | | | | *1,287* | *91.28%* | *1,217* | *90.68%* | *1,879* | *45.21%* |
| **Intergenic** | Intergenic | | | 123 | 8.72% | 125 | 9.31% | 2,277 | 54.79% |
| *Subtotal (Intergenic)* | | | | *123* | *8.72%* | *125* | *9.31%* | *2,277* | *54.79%* |
| **Sum** | | | | 1,410 | 100.00% | 1,342 | 100.00% | 4,156 | 100.00% |

Using the X-square test, the proportion of intergenic expression candidates relative to all candidate pool in wild-type (9%) and *ibm2* (9%) is shown to be significantly lower than the expected insertion distribution estimated from of all annotated loci (42%; Figure 5.7 A), whereas the percentage of intergenic expression candidates increased dramatically to 55% in *ddm1,* which is significantly higher than the proportion in all three situations mentioned above (Figure 5.7 A). For the percentage of 'gene unit' expression candidates relative to all genic expression candidates (Figure 5.7 B), this proportion in wild-type (82%) and *ibm2* (81%) is significantly higher than the expected proportion estimated from all annotated loci (27%), while this 'gene-unit' proportion in *ddm1* (55%) is significantly skewed from all of the other three situations (Figure 5.7 B). The comparison between the expected and observed ratio of expression candidates co-localized with either expressed or non-expressed genes revealed a significant bias towards expressed genes in all genotypes (Figure 5.7 C). However, the proportion of those co-localized with inactive genes in *ddm1* was twice as high as that found in the wild-type and *ibm2* datasets. Furthermore, compared with wild-type and *ibm2*,

deficiency in *DDM1* also revealed that there was a significant 5% increase in full-length expression candidates (Figure 5.7 D) and a 7% increase in un-trackable loci (Figure 5.7 E). In wild-type and *ibm2* datasets, 82.1% and 83.8% of the full-length expression candidates co-localized with genes, respectively, whereas in the *ddm1* dataset, only 51% of the full-length candidates were with genes, leaving 48.7% of the full-length candidates in the intergenic region. Likewise, 79.2% and 88% of the un-trackable expression candidates in wild-type and *ibm2,* respectively, were co-localized with genes. In contrast, only 27% of the un-trackable candidates in *ddm1* co-localized with genes, while 73% of the un-trackable expression candidates in *ddm1* were in the intergenic region.

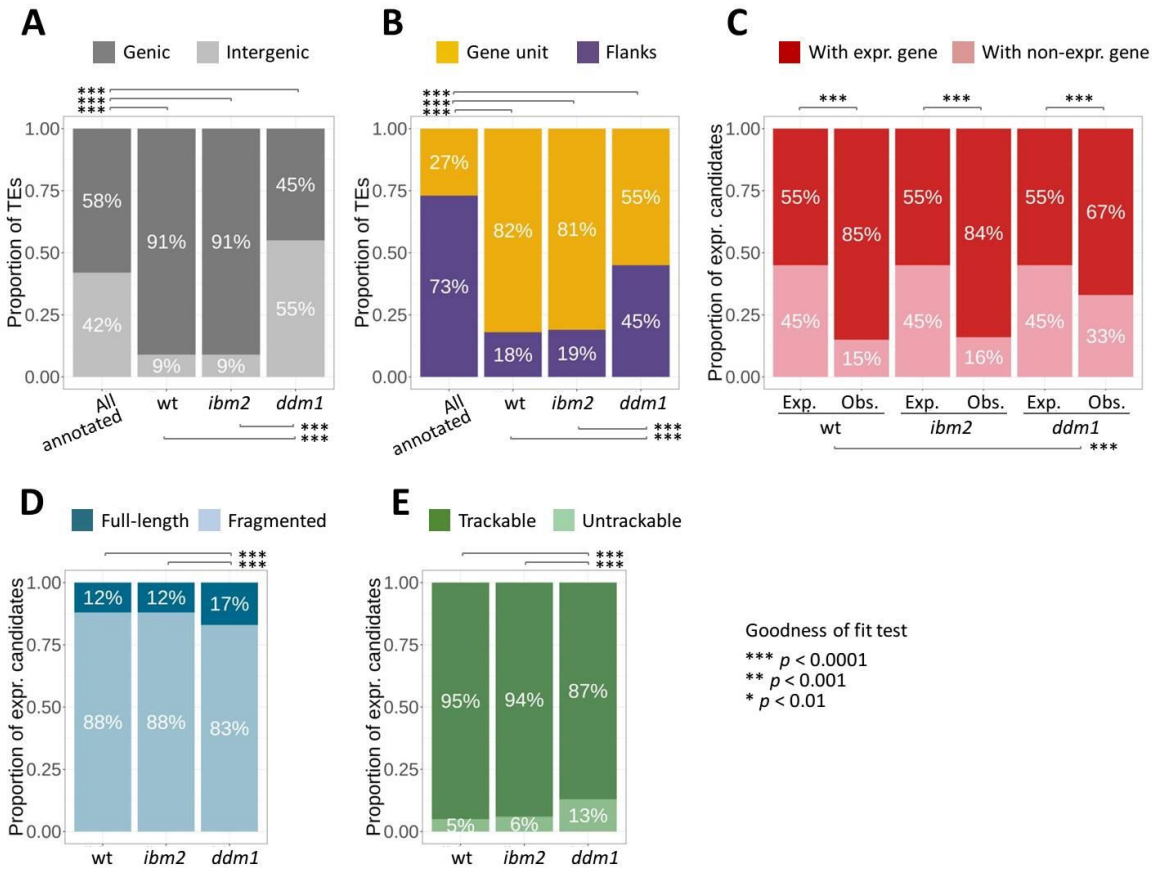



**Figure 5.7 Characteristics of *A. thaliana* expression candidates in terms of location, integrity and distinctness.**

**(A)** Categorization of annotated TEs and expression candidates by genic/intergenic regions. **(B)** Categorization of annotated genic TEs and genic expression candidates by location relative to genes. **(C)** Classification of genic expression candidates by the transcriptional activity of co-localized genes and statistical comparison between the expected and observed values. **(D-E)** Categorization of all expression candidates by integrity (D) and distinctness (E). The goodness of fit test was performed pair-wisely. All the comparisons reached $p < 0.01$ were labelled. Levels of statistical significance were as indicated. Exp., expected; Obs., observed.

### 5.4.4 Identification of potential origins of *Arabidopsis* autonomous retrotransposon transcripts

Among the 5,962 LTR-type transposable elements (LTR-TEs) annotated in TAIR10 *A. thaliana* reference genome, 466 are full-length TEs, of which 190 retain LTRs at both ends (Table 5.4). By use of the same filtering approach described in section 3.3.8, only 41 and 47 intact LTR-TEs were identified in the candidate pools of wild-type and *ibm2*, respectively. Further investigation revealed three loci for each of the two genotypes with over 90% breadth of coverage across the INT domain. In contrast, 350 full-length LTR-TE expression candidates were found in *ddm1*. This includes 173 intact loci flanked with LTRs, of which 87 loci showed over 90% coverage throughout the INT domain. The Venn diagram in Figure 5.8 A demonstrated that these 87 loci include the three individual elements found in wild-type and *ibm2* as well. A further categorization of the 87 autonomous candidates in *ddm1* revealed that over half were localised in the intergenic region (Figure 5.8 B). The potentially autonomous expression candidates in the genic region were split equally into two groups, co-localizing with expressed genes or non-expressed genes, each predominantly comprised of candidates in the flanking regions (Figure 5.8 C). These 87 loci were derived from 46 LTR-TE families, and none of these families obtained more than ten autonomous LTR-TE expression candidates (Figure 5.9).

**Table 5.4 Number of selected *A. thaliana* TEs at each stage in the workflow of collecting potential origins of autonomous Type I LTR-TE transcripts.**

| TE subsets | Treatments | # Selected TEs | | |
|---|---|---|---|---|
| | | Full-length | Full-length with LTRs | >90% INT coverage |
| **Annotated TEs** | | 466 → | 190 | - |
| **Expr. candidates** | Wild-type | 41 → | 14 → | 3 |
| | *ibm2* | 47 → | 18 → | 3 |
| | *ddm1* | 350 → | 173 → | 87 |

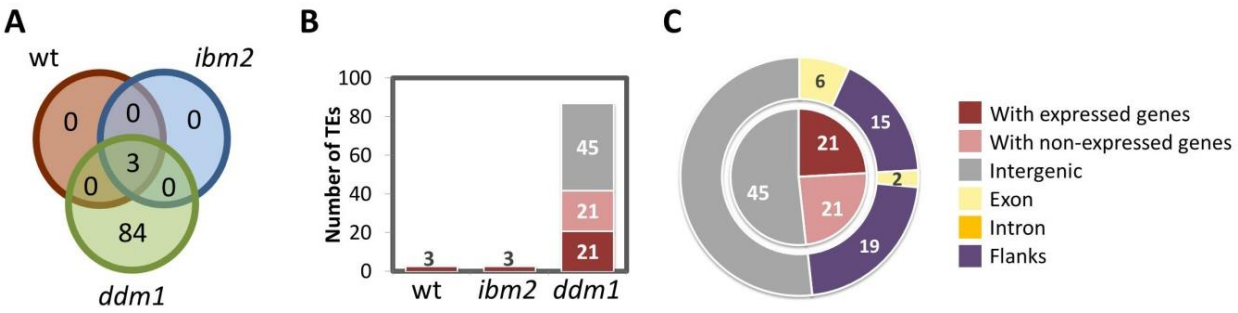



**Figure 5.8 Comparison of potential autonomous LTR-TE expression candidates in *A. thaliana* across genotypes**

**(A)** The comparison of the three sets of potential autonomous LTR-TE candidates was illustrated by the Venn diagram. **(B-C)** Categorization by location. As the three loci found in wild-type and ibm2 were all included in the ddm1 collection, the hierarchical categorization of all 87 loci is summarized in Figure 5.9.

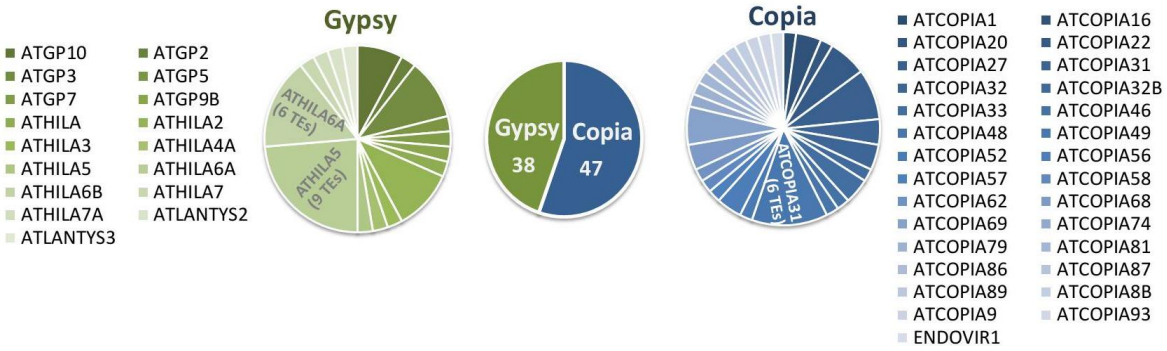



**Figure 5.9 Categorization of the potential autonomous LTR-TE expression candidates in *A. thaliana* by family**

The 87 loci were grouped into either Copia or Gypsy superfamily (centre), following with further categorization by family shown in the pie graph on both sides. Family obtaining more than five autonomous candidates were indicated on the corresponding slices.

LINE elements are the second largest class of type I retrotransposon after LTR-TE in *A. thaliana*, and are represented by 1,447 annotated loci derived from 12 TE families. However, only 110 loci are full length, of which 49 belong to LINE families that appear to be intact and capable of retrotransposition as defined by having a complete reverse transcriptase (RT) domain in the canonical sequences (Table 5.5). Among the 49 competent LINE loci, four were potentially expressed in wild-type and *ibm2,* and 38 were found in *ddm1*. Nonetheless, none of the competent LINE candidates in the former two genotypes showed the breadth of coverage exceeding 90% of the element body, whereas 17 of the 38 loci in *ddm1* did. Focusing on the 17 potential origins of autonomous LINE transcripts, about one-

third (6 loci) of them were intergenic, and the majority of the remainder (10 out of 11) were located in flanking regions (Figure 5.10 A). Fourteen of the 17 autonomous LINE candidates in *ddm1* came from either L1-ATLINE1_6 (7 loci) or L1-TA11 (7 loci) families, while the remaining three were from L1-ATLINE1_2 and L1-ATLINE1_4 (Figure 5.10 B).

**Table 5.5 Number of selected *A. thaliana* TEs at each stage in the workflow of collecting potential origins of autonomous Type I LINE transcripts.**

The competent family denotes those retaining intact reverse transcriptase (RT) domain with putative active sites in the canonical sequence.

| TE subsets | Treatments | # Selected TEs | | |
|---|---|---|---|---|
| | | Full-length | Full-length TE of competent family | >90% coverage |
| **Annotated TEs** | | 110 → | 49 | - |
| **Expr. candidates** | Wild-type | 8 → | 4 → | 0 |
| | *ibm2* | 8 → | 4 → | 0 |
| | *ddm1* | 45 → | 38 → | 17 |

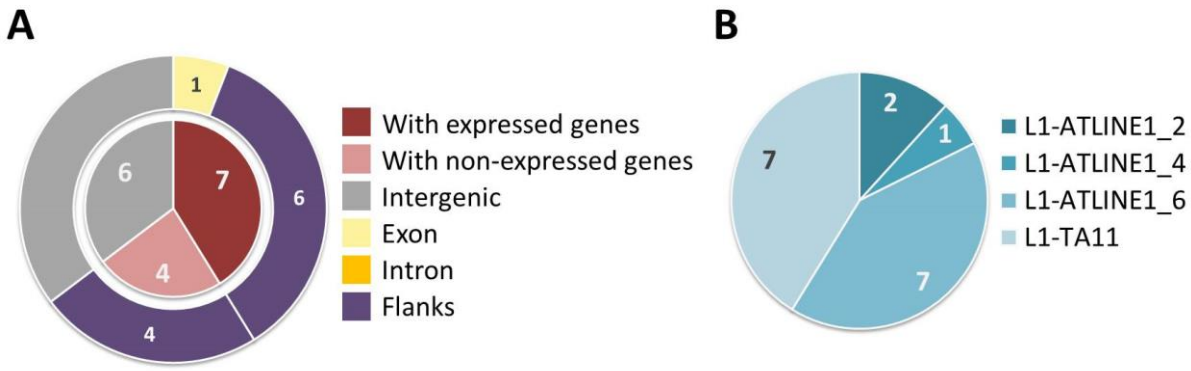



**Figure 5.10 Categorization of the potential autonomous LINE expression candidates in *A. thaliana* by location and family**

The 17 loci found in ddm1 were grouped by (A) location and (B) family. The digits in each slice or segment denote the number of TE loci in each category.

### 5.4.5 Collection of *D. melanogaster* expression candidates using the pipeline

To explore whether the characteristics of potentially expressed TEs in *V. vinifera* and *A. thaliana* were conserved across kingdoms, transcriptome data from the ALS model in *Drosophila melanogaster* (Krug et al. 2017) were extracted and analyzed in the same way. In agreement with Krug et al. (2017), the mapping statistics (Table 5.6) showed that roughly 56 to 70% of sequencing reads were mapped to the *D. melanogaster* dm3 reference genome. Respectively, 50.42%, 44.73%, and 46.33% of the 74,448 annotated *Drosophila* TEs were found potentially expressed in the control, glial, and neuronal ALS models, respectively (Figure 5.11 A-C), unlike the small proportion of expression candidates seen in grapevine (1.5-3%) and *Arabidopsis* (4-14%, Table 5.7). In wild-type control flies, 30.93% of transcribed TEs were sparsely and non-uniformly covered by reads thus fell under the threshold, accompanying with 18.65% of TE loci that had no reads mapped (Figure 5.11 A). This pattern was also found in transgenic flies, which exhibited ectopic expression of human TDP-43 (hTDP-43) in *Drosophila*'s glial and neuronal cells (Figure 5.11 B-C). Irrespective of whether there was an ectopic expression of hTDP-43 in *Drosophila*'s glial and neuronal cells, about two-thirds of the expression candidates were un-trackable candidates in which there was a lack of unique-mapping reads (Figure 5.11 D-F), compared to approximately one-third of TE expression candidates that were un-trackable in grapevine and *Arabidopsis* (Table 5.7).

**Table 5.6 Mapping statistics for RNA-seq analysis of *Drosophila* dataset of Krug *et al.*.**

| Sequenced libraries | | Sequenced reads | | Adaptor removal | | Filter tRNA/rRNA | | Mapped reads | |
|---|---|---|---|---|---|---|---|---|---|
| Genotypes | Replicates | | | | | | | | |
| hTDP-43 / + (Control) | a | 150,105,062 | 100% | 127,807,740 | 85.15% | 108,504,090 | 72.29% | 92,732,212 | 61.78% |
| | b | 126,254,836 | 100% | 107,448,436 | 85.10% | 87,625,822 | 69.40% | 73,269,755 | 58.03% |
| *Repo* > hTDP-43 (Glial hTDP-43) | a | 152,058,342 | 100% | 131,254,824 | 86.32% | 118,460,876 | 77.90% | 102,463,100 | 67.38% |
| | b | 150,691,824 | 100% | 122,225,190 | 81.11% | 104,934,688 | 69.64% | 85,688,996 | 56.86% |
| *ELAV* > hTDP-43 (Neuronal hTDP-43) | a | 151,245,166 | 100% | 129,402,444 | 85.56% | 111,804,552 | 73.92% | 96,670,858 | 63.92% |
| | b | 152,632,590 | 100% | 131,678,642 | 86.27% | 122,121,516 | 80.01% | 107,377,195 | 70.35% |

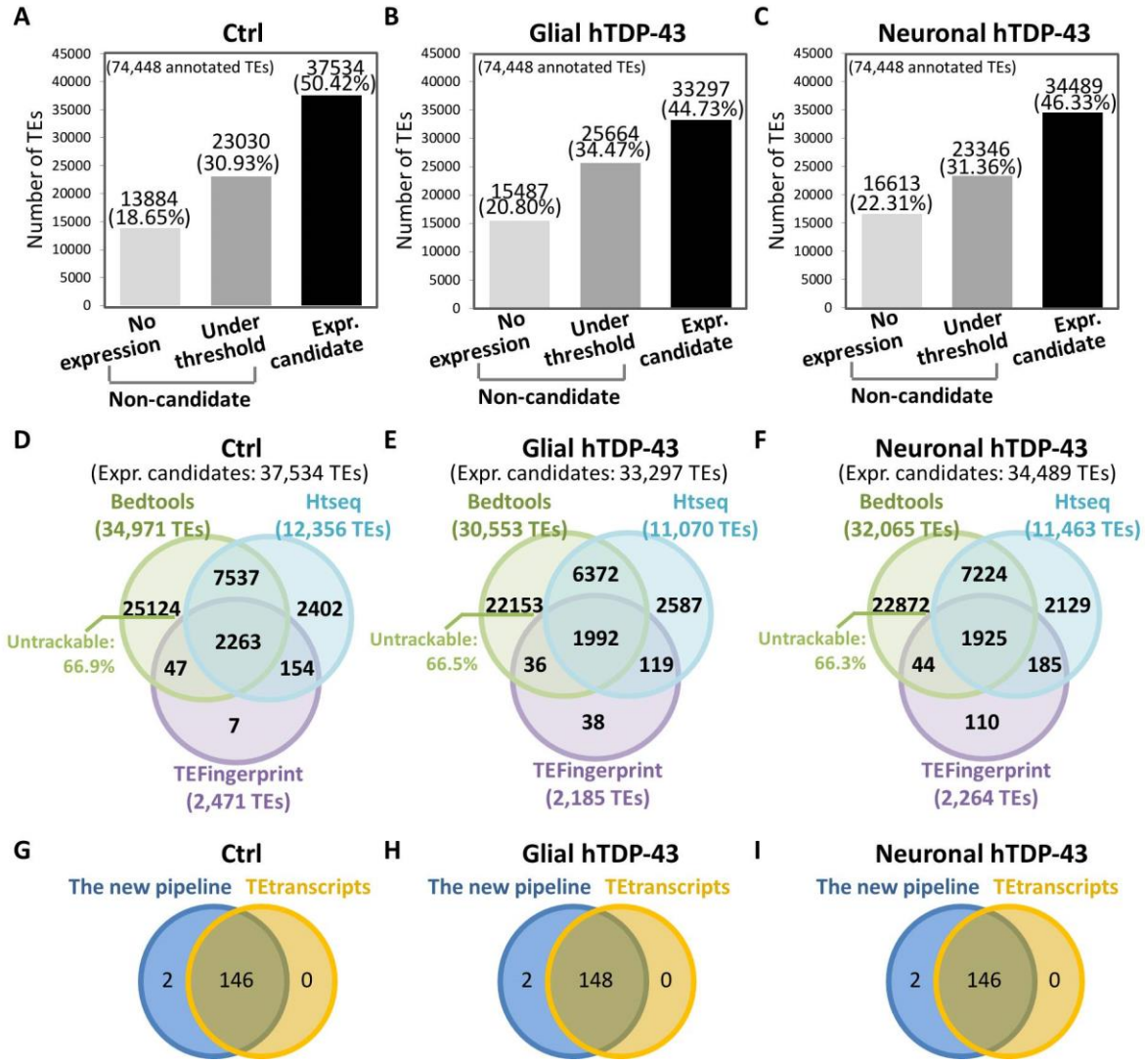



**Figure 5.11 Expression candidates of the *Drosophila* TDP-43 model of ALS (Krug *et al*. 2017) identified by the pipeline**

**(A)-(C)** All annotated TEs were categorized by transcriptional activity and illustrated according to treatment as indicated. **(D)-(F)** The expression candidates of each treatment was a pool of potentially expressed TEs collected by the three sub-pipelines. **(G)-(I)** The expression candidates were grouped by family to identify transcriptionally active TE families, which were then further compared with the active families captured by TEtransctipts.

Expressed TE families retrieved from our pipeline were compared with the results obtained from the commonly used software TEtranscripts. The data collected from these two methods overlapped TEtranscripts fully overlapped the pool of families identified by our pipeline (Figure 5.11 G-I). TE families uniquely found by our pipeline are listed in Table 5.8. The family UHU only contained one expression candidate having a low number of unique-mapping reads that nearly passed the threshold in the htseq-count sub-pipeline yet failed in the TEtranscripts-based method. The TE families 5S_DM and NOF_FB were comprised of expression candidates sharing the majority of sequencing reads with genes and other TE families (Figure 5.12). As a result, these reads might have been predominantly

assigned to genes, or, during estimation of relative abundance for each individual TE, the Expectation Maximization algorithm implemented in TEtranscripts might favour TE loci of other families sharing sequencing reads with 5S_DM and NOF_FB families. Although these TEs may not be the origins of these reads, they could be targeted or modulated by the epigenetic regulatory system together with the associated genes and TE families.

**Table 5.7 Comparison of the proportion of expression candidates categorized by read specificity.**

| Species | Annotated TEs | | Expression candidates | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | TE Categories | Control (T=0) | | Mock | | Yeast | | Botrytis | |
| *V. vinifera* | 223,411 | 100% | Trackable candidates | 1,140 | 0.51% | 1,548 | 0.69% | 1,250 | 0.56% | 1,120 | 0.50% |
| | | | Untrackable candidates | 2,558 | 1.15% | 3,976 | 1.78% | 4,281 | 1.91% | 4,051 | 1.81% |
| | | | Sum | 3,698 | 1.66% | 5,524 | 2.47% | 5,531 | 2.47% | 5,171 | 2.31% |
| | | | TE Categories | Wild-type (*Col*) | | *ibm2* | | *ddm1* | | | |
| *A. thaliana* | 31,189 | 100% | Trackable candidates | 1,338 | 4.29% | 1,267 | 4.06% | 3,619 | 11.60% | | |
| | | | Untrackable candidates | 72 | 0.23% | 75 | 0.24% | 537 | 1.72% | | |
| | | | Sum | 1,410 | 4.52% | 1,342 | 4.30% | 4,156 | 13.32% | | |
| | | | TE Categories | Control (hTDP-43/+) | | Glial hTDP-43 (*Repo* > hTDP-43) | | Neuronal hTDP-43 (*ELVA* > hTDP-43) | | | |
| *D. melanogaster* | 74,448 | 100% | Trackable candidates | 25,124 | 33.75% | 22,153 | 29.76% | 22,872 | 30.72% | | |
| | | | Untrackable candidates | 12,410 | 16.67% | 11,144 | 14.97% | 11,617 | 15.61% | | |
| | | | Sum | 37,534 | 50.42% | 33,297 | 44.73% | 34,489 | 46.33% | | |

**Table 5.8 Expressed *Drosophila* TE families uniquely found from the collections of expression candidates.**

| Genotype | TE families | # expr. candidates | Possible reasons for the exclusion from TEtranscripts approach |
|---|---|---|---|
| Control (hTDP-43/+) | NOF_FB | 2 | A high proportion of NOF_FB-related reads also mapped to genes and TE individuals of other families. |
| | 5S_DM | 178 | A high proportion of 5S_DM-related reads also mapped to genes and TE individuals of other families. |
| Glial hTDP-43 (*Repo* > hTDP-43) | NOF_FB | 1 | A high proportion of NOF_FB-related reads also mapped to genes and TE individuals of other families. |
| | UHU | 1 | Expression candidates of this family nearly passed the threshold of the htseq-count sub-pipeline yet fell in the TEtranscripts approach. |
| Neuronal hTDP-43 (*ELVA* > hTDP-43) | NOF_FB | 11 | A high proportion of NOF_FB-related reads also mapped to genes and TE individuals of other families. |
| | 5S_DM | 250 | A high proportion of 5S_DM -related reads also mapped to genes and TE individuals of other families. |

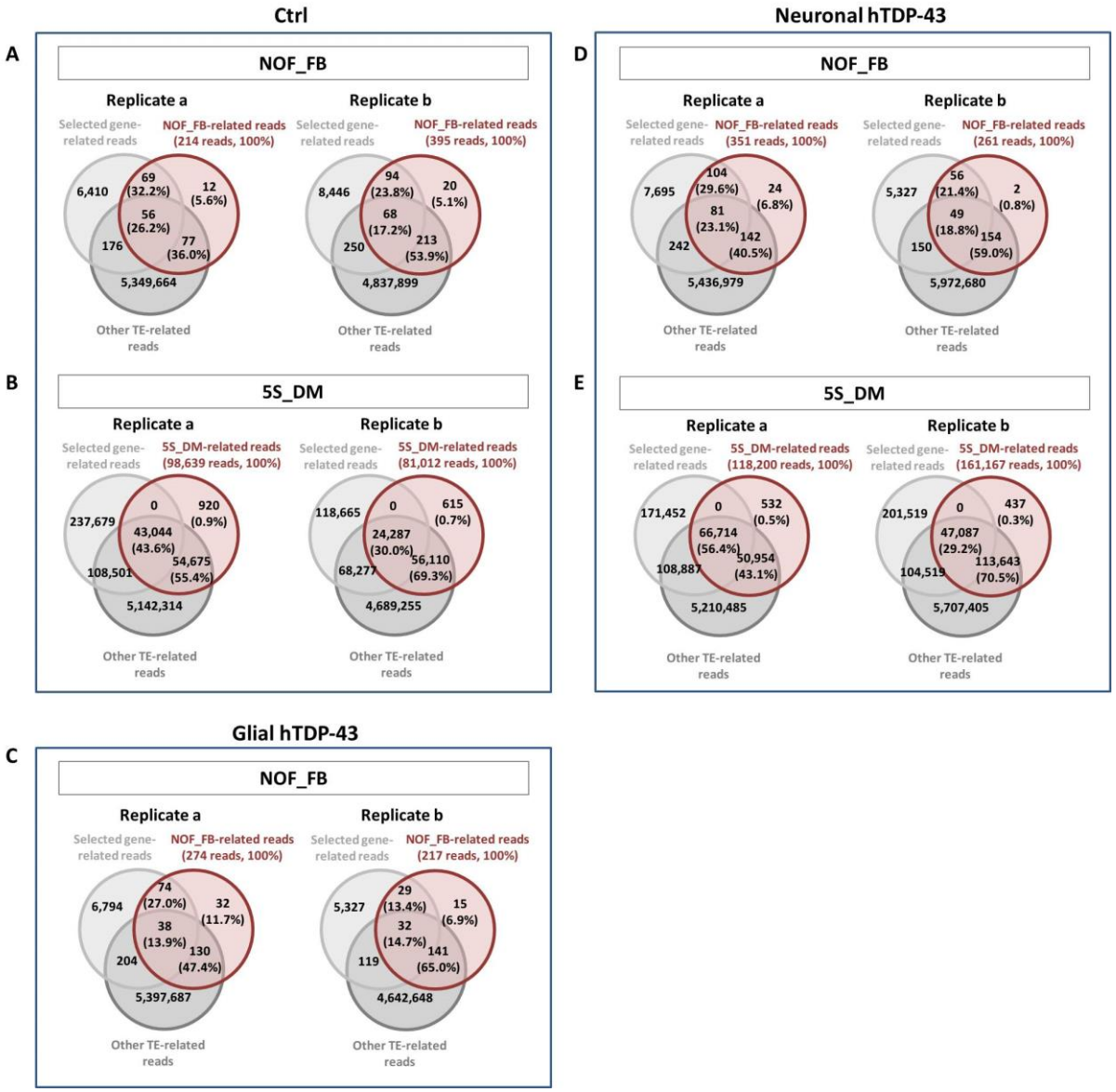



**Figure 5.12 Sequencing reads related to *Drosophila* TE families excluded in the TEtranscripts approach were also shared by other TE families and genes.**

**(A-B)** In control, the TE family NOF_FB (A) and 5S_DM (B) respectively shared about 95% and 99% of reads with genes and other TEs. **(C-E)** This phenomenon was also observed in the NOF_FB in the glial model (C) and the neuronal model (D), as well as 5S_DM in the neuronal model (E).

To investigate whether the ectopic expression of hTDP-43 would lead to a shift in the population of expression candidates, the three sets of expression candidates from the three genotypes were compared and illustrated in Figure 5.13. Although 3,501, 1,612, and 1,493 loci were, respectively, unique to the control, glial, and neuronal models, the central overlapping area revealed a tremendous number (29,096 loci) of expression candidates appeared in all genotypes (Figure 5.13 A). Analysing at the family level showed that, among the total 159 TE families annotated in the reference

genome, almost all transcriptionally active TE families (148 of 150 active families) in this dataset were conserved across the three genotypes (Figure 5.13 B).

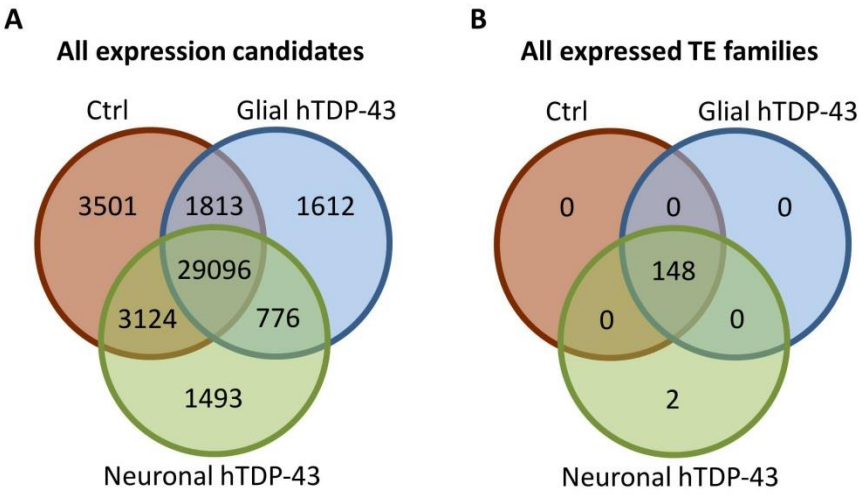



**Figure 5.13 Comparison of *Drosophila* expression candidates and active TE families among the three genotypes.**

Comparisons of different sets of **(A)** expression candidates and **(B)** active TE families from the indicated genotypes were illustrated by Venn diagrams, where the overlapping areas denote expression candidates appeared in more than one genotype and the non-overlapping areas denote those exclusively found in specific genotype.

### 5.4.6 The integrity of the *D. melanogaster* expression candidates

Based on the reference genome, about 93% of all 74,488 annotated TEs are fragmented remnants of intact elements, with only 7.23% retaining over 90% coverage of the canonical sequences (Figure 5.14 A). In control flies, full-length TEs only contributed to 5.91% of the expression candidate pool comprising 37,537 TEs (Figure 5.14 B). As the fraction of full-length candidates in the glial hTDP-43 model increased slightly to 6.21 % (Figure 5.14 C), the full-length proportion in the neuronal model (Figure 5.14 D) was similar to that in control flies.

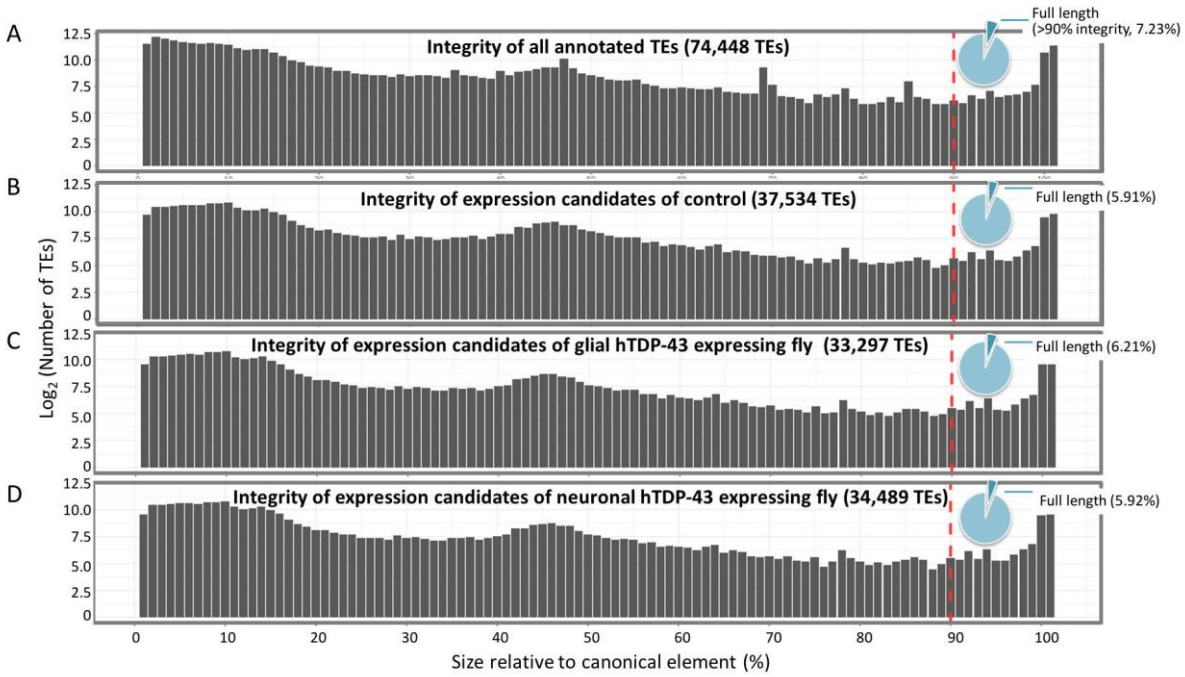



**Figure 5.14 Integrity of annotated TEs and expression candidates of *Drosophila* TDP-43 model of ALS.**

The length of each individual TE locus was compared with the size of the corresponding canonical element. TEs with over 90% integrity were denoted as full length. **(A)** All annotated TEs (74,448 TEs). **(B)-(D)** Expression candidates of control (B), glial hTDP-43 (C) and neuronal hTDP-43 flies.

### 5.4.7 Hierarchical classifications of *D. melanogaster* expression candidates by location, integrity, and distinctness

The location preference of expression candidates in *Drosophila* is also different from what we have observed in both grapevine and *Arabidopsis*. About half of the annotated TEs in fruit fly locate in genic regions (Figure 5.15 A, Table 5.9), where intronic TEs are overrepresented. The proportion of genic expression candidates remained similar in the control model (Figure 5.15 B); however, as opposed to the location bias towards expressed genes in grapevine and *Arabidopsis*, the majority of genic candidates co-localized with non-expressed genes in control flies. Given the strong prevalence of intronic insertions observed from all annotated genic TEs, the majority of the genic expression candidates were found in introns as well. The distribution of expression candidates in terms of insertion location in the glial and neuronal hTDP-43 models (Figure 5.15 C, D) remained consistent with the control model, except that, in the glial model, the fraction of expression candidates co-localizing with expressed genes increased by 8.5 % (Table 5.10).

**Table 5.9 Hierarchical categorization of *D. melanogaster* annotated TEs by location and integrity.**

Numbers of TEs (#TE) in black colour sum up to 74,448 annotated TEs, while the corresponding percentages (Perc.) in black colour sum up to 100.00%. Subtotals of indicated categories were denoted in grey.

| Region | Location | | All annotated TEs | |
|---|---|---|---|---|
| | | | #TE | Perc. |
| **Genic** | Gene unit | Exon | 1,156 | 1.55% |
| | | Intron | 34,171 | 45.90% |
| | *Subtotal (Gene unit)* | | *35,327* | *47.45%* |
| | Flanks | N-Flank | 1,337 | 1.80% |
| | | N&C-Flank | 590 | 0.79% |
| | | C-Flank | 1,384 | 1.86% |
| | *Subtotal (Flanks)* | | *3,311* | *4.45%* |
| *Subtotal (Genic)* | | | *38,638* | *51.90%* |
| **Intergenic** | | | 35,810 | 48.10% |
| *Subtotal (Intergenic)* | | | *35,810* | *48.10%* |
| **Sum** | | | 74,448 | 100.00% |

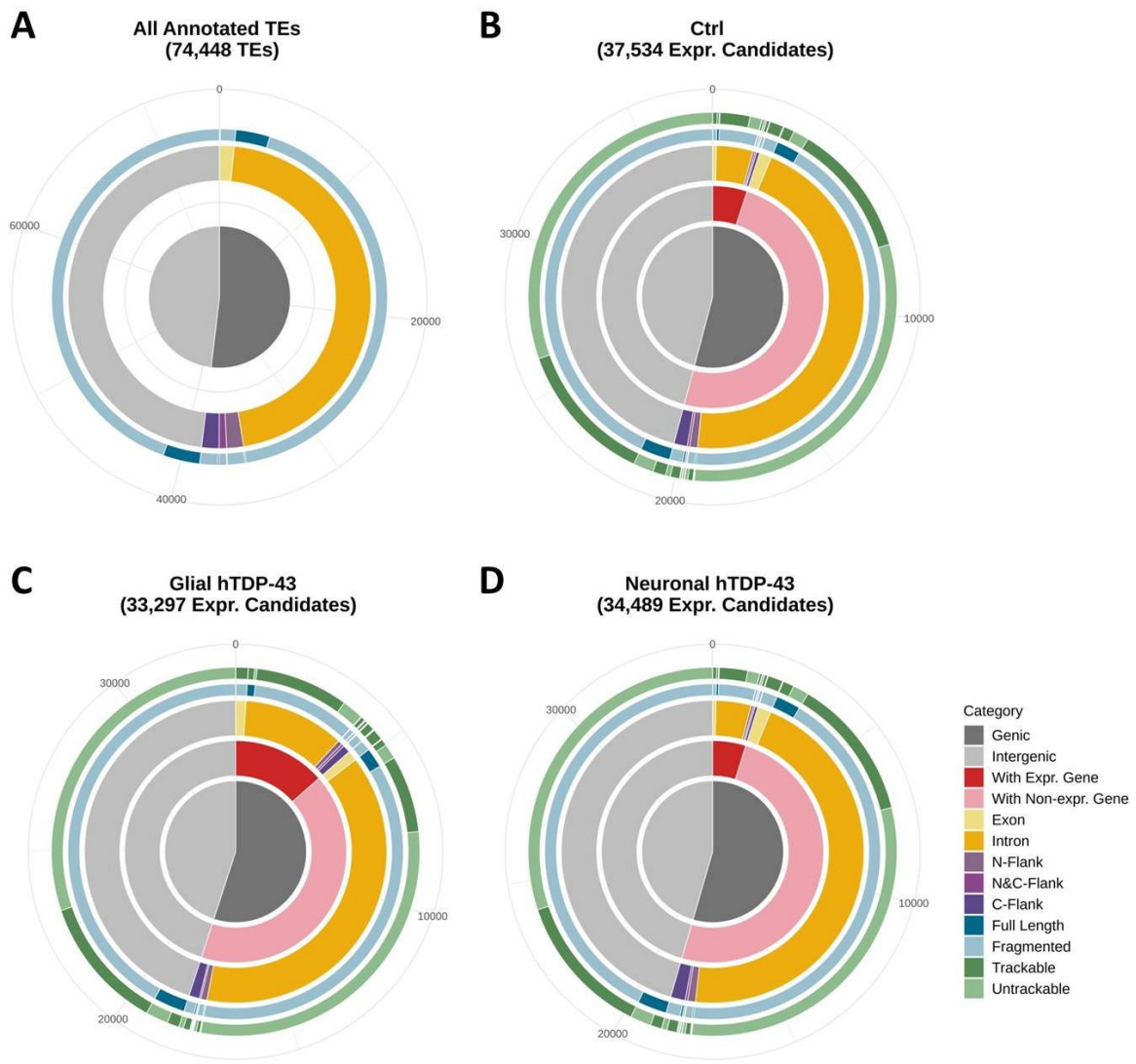



**Figure 5.15 Hierarchical classifications of *D. melanogaster* expression candidates by location, integrity, and distinctness.**

**(A)** TEs overlapping with exon, intron, or 2kb upstream (N-terminus) or downstream (C-terminus) of a gene were denoted as genic TEs; otherwise, they were grouped as intergenic TEs. **(B)** All annotated TEs were categorized hierarchically by region (centre), location (internal layer) and integrity (outer-most layer). **(C-E)** Expression candidates of each genotype were categorized in the order of region (centre), the transcriptional activity of co-localized genes (2$^{nd}$ layer), location (3$^{rd}$ layer), integrity (4$^{th}$ layer), and the presence/absence of unique-mapping reads (outer-most layer).

**Table 5.10 Hierarchical categorization of *D. melanogaster* expression candidates by location and integrity.**

Numbers of TEs (#TE) in black colour sum up to 74,448 annotated TEs, while the corresponding percentages (Perc.) in black colour sum up to 100.00%. Subtotals of indicated categories were denoted in grey.

| Region | Gene activity | Location | | Control | | Glial hTDP-43 | | Neuronal hTDP-43 | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | #TE | Perc. | #TE | Perc. | #TE | Perc. |
| **Genic** | With expr. Gene | Gene unit | Exon | 151 | 0.40% | 369 | 1.11% | 124 | 0.36% |
| | | | Intron | 1,470 | 3.92% | 3,604 | 10.82% | 1,296 | 3.76% |
| | | *Subtotal (Gene unit)* | | *1,621* | *4.32%* | *3,973* | *11.93%* | *1,420* | *4.12%* |
| | | Flanks | N-Flank | 87 | 0.23% | 163 | 0.49% | 87 | 0.25% |
| | | | N&C-Flank | 60 | 0.16% | 92 | 0.28% | 53 | 0.15% |
| | | | C-Flank | 118 | 0.31% | 273 | 0.82% | 106 | 0.31% |
| | | *Subtotal (Flanks)* | | *265* | *0.71%* | *528* | *1.59%* | *246* | *0.71%* |
| | *Subtotal (With expr. gene)* | | | *1,886* | *5.03%* | *4,501* | *13.52%* | *1,666* | *4.83%* |
| | With non-expr. Gene | Gene unit | Exon | 495 | 1.32% | 357 | 1.07% | 498 | 1.44% |
| | | | Intron | 16,982 | 45.24% | 12,805 | 38.46% | 15,708 | 45.55% |
| | | *Subtotal (Gene unit)* | | *17,477* | *46.56%* | *13,162* | *39.53%* | *16,206* | *46.99%* |
| | | Flanks | N-Flank | 320 | 0.85% | 214 | 0.64% | 290 | 0.84% |
| | | | N&C-Flank | 93 | 0.25% | 57 | 0.17% | 87 | 0.25% |
| | | | C-Flank | 538 | 1.43% | 376 | 1.13% | 539 | 1.56% |
| | | *Subtotal (Flanks)* | | *951* | *2.53%* | *647* | *1.94%* | *916* | *2.66%* |
| | *Subtotal (With non-expr. gene)* | | | *18,428* | *49.10%* | *13,809* | *41.47%* | *17,122* | *49.65%* |
| *Subtotal (Genic)* | | | | *20,314* | *54.12%* | *18,310* | *54.99%* | *18,788* | *54.48%* |
| **Intergenic** | Intergenic | | | 17,220 | 45.88% | 14,987 | 45.01% | 15,701 | 45.53% |
| *Subtotal (Intergenic)* | | | | *17,220* | *45.88%* | *14,987* | *45.01%* | *15,701* | *45.53%* |
| **Sum** | | | | 37,534 | 100.00% | 33,297 | 100.00% | 4,156 | 100.00% |

Although the location bias seemed trivial, a X-square test showed that, in all genotypes, the genic fractions of expression candidates were significantly higher ($p = 2.149e-12$, $p < 2.2e-16$, and $p = 2.475e-15$ in control, glial, and neuronal models, respectively) than the expected percentage based on the distribution of all annotated TEs (Figure 5.16 A). Also, compared to the expected distribution of genic TEs estimated from all annotated loci (91%), the proportion of expression candidates present in gene units relative to all expression candidates in the genic region was increased by 3% in all genotypes ($p < 2.2e-16$, Figure 5.16 B). Furthermore, for genic expression candidates, the observed distribution bias towards non-expressed genes was significantly stronger by 7% in control ($p < 2.2e-16$), 2% in glial ($p = 5.263e-08$), and 6% in neuronal ($p < 2.2e-16$) models than the expected distribution estimated by the ratio of expressed to non-expressed genes (Figure 5.16 C). Intriguingly, the proportion of expression candidates co-localized with expressed genes in the glial model was remarkably larger than control ($p < 2.2e-16$) and neuronal ($p < 2.2e-16$) models, seemingly in

concordance with the ratio of expressed and non-expressed genes. Our analysis on TE integrity (Figure 5.16 D) and whether loci were distinguishable (Figure 5.16 E), by contrast, were not found to be significantly different among genotypes.

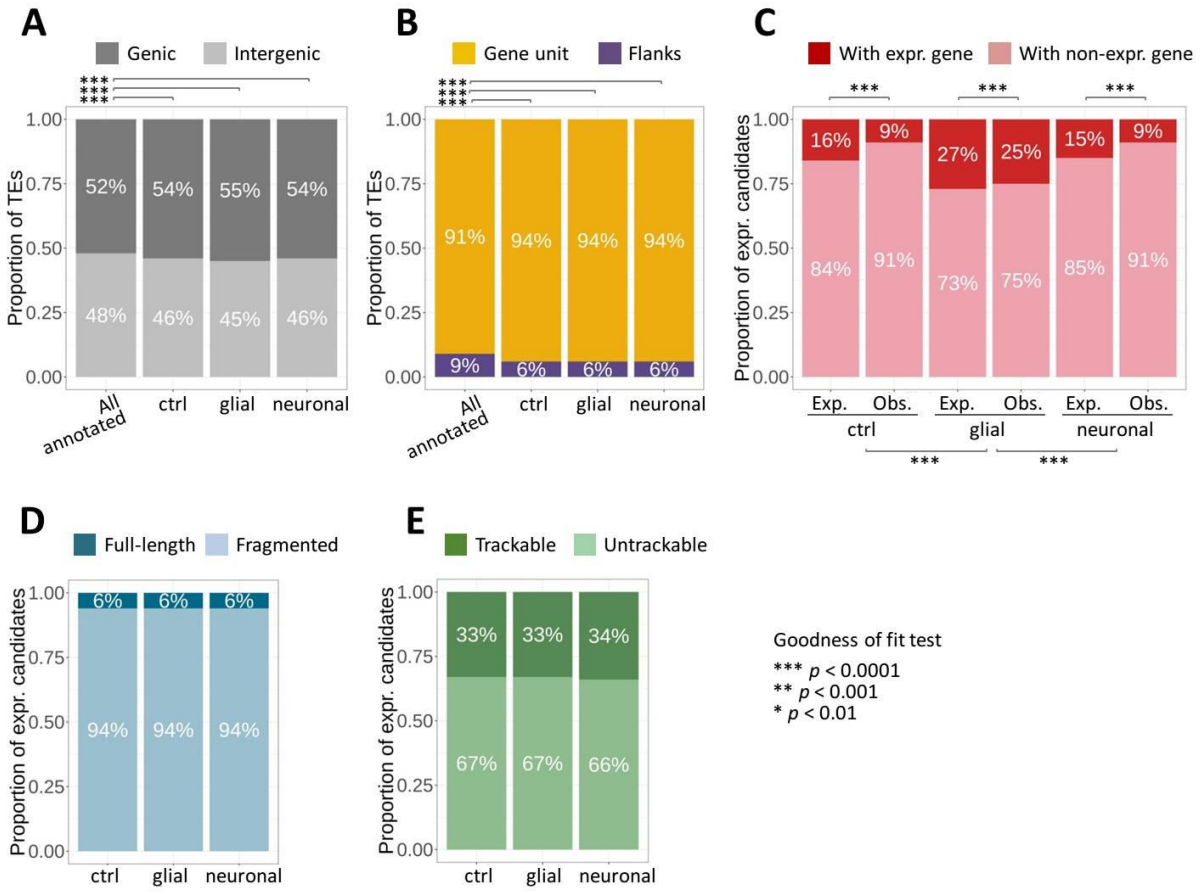



**Figure 5.16 Characteristics of *D. melanogaster* expression candidates in terms of location, integrity and distinctness.**

**(A)** Categorization of annotated TEs and expression candidates by genic/intergenic regions. **(B)** Categorization of annotated genic TEs and genic expression candidates by location relative to genes. **(C)** Classification of genic expression candidates by the transcriptional activity of co-localized genes and statistical comparison between the expected and observed values. **(D-E)** Categorization of all expression candidates by integrity (D) and distinctness (E). The goodness of fit test was performed pair-wisely. All the comparisons reached $p < 0.01$ were labelled. Levels of statistical significance were as indicated. Exp., expected; Obs., observed.

### 5.4.8 Identification of potential origins of *Drosophila* autonomous retrotransposon transcripts

Having contributed to over one-third of TE loci (27,929 TEs) deposited in the *D. melanogaster* dm3 reference genome, the LTR-type retrotransposon (LTR-TE) includes 955 full-length TEs, of which 332 retain LTRs at both ends (Table 5.11). By use of the same filtering approach for the analysis in grapevine and *Arabidopsis*, 773, 744 and 777 full-length LTR-TE expression candidates were identified in the control, glial and neuronal hTDP-43 models, respectively. Remarkably, nearly 90% of these TE loci were found to be expression candidates that demonstrated > 90% breadth of coverage of RNA-seq reads over the INT domain (Table 5.11). As expected, a large proportion of these competent candidates were shared across the three genotypes at both individual TE (Figure 5.17 A) and TE-family level (Figure 5.17 B). Interestingly, the LTR family TIRANT was found to contain 17 loci that exhibited potential transcription across the entire element, specifically in response to the ectopic hTDP-43 expression in neuronal cells. Although there were also TE families ACCORD2 and FROGGER that were unique to healthy flies and the glial model, respectively, each of which showed evidence of only one potential origin of autonomous LTR-TE transcripts. The location distribution showed that 62%, 63% and 65% of these autonomous expression candidates were co-localized with genes in the control, glial model and neuronal model, respectively (Figure 5.18). Although the proportion of autonomous LTR-TE expression candidates in the genic region was similar between these genotypes, the autonomous candidates from the glial ALS model show a higher tendency to co-localize with expressed genes than the other two genotypes. In the glial ALS model, 27% of the autonomous LTR-TE expression candidates co-localized with expressed genes (Figure 5.18 B), while in the control and neuronal model, 8.5% and 8.2% of the autonomous LTR-TE expression candidates were with expressed genes, respectively (Figure 5.18 A, C). The majority of these autonomous expression candidates were originated from the Gypsy superfamily, in which 25 Gypsy families contributed 190 autonomous LTR-TE candidates (Figure 5.19); yet the ROO element, with 62 autonomous candidates belongs to the Pao superfamily, was the most overrepresented family followed by Copia having 32 autonomous candidates.

**Table 5.11 Number of selected *D. melanogaster* TEs at each stage in the workflow of collecting potential origins of autonomous Type I LTR-TE transcripts.**

| TE subsets | Treatments | # Selected TEs | | |
|---|---|---|---|---|
| | | Full-length | Full-length with LTRs | >90% INT coverage |
| **Annotated TEs** | | 955 → | 332 | - |
| **Expr. candidates** | Control | 773 → | 316 → | 284 |
| | Glial hTDP-43 | 744 → | 315 → | 281 |
| | Neuronal hTDP-43 | 777 → | 331 → | 292 |

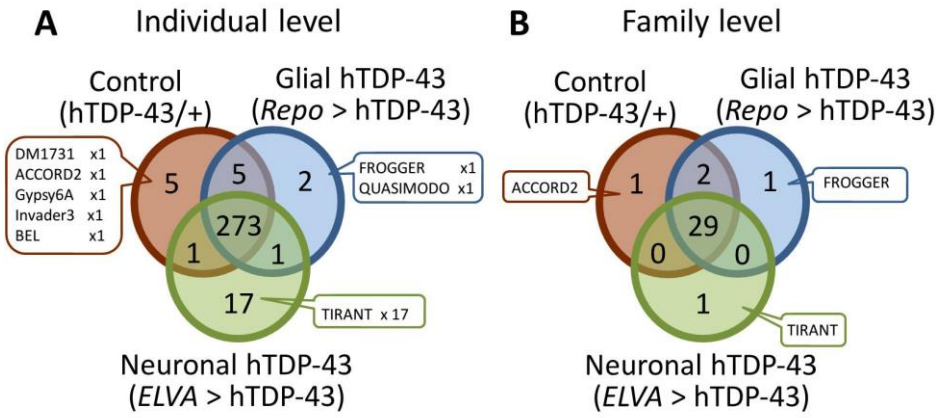



**Figure 5.17 Comparison of potential autonomous LTR-TE expression candidates in *D. melanogaster* across genotypes**

The comparison of the three sets of potential autonomous LTR-TE candidates was illustrated by the Venn diagram at **(A)** individual and **(B)** family levels.

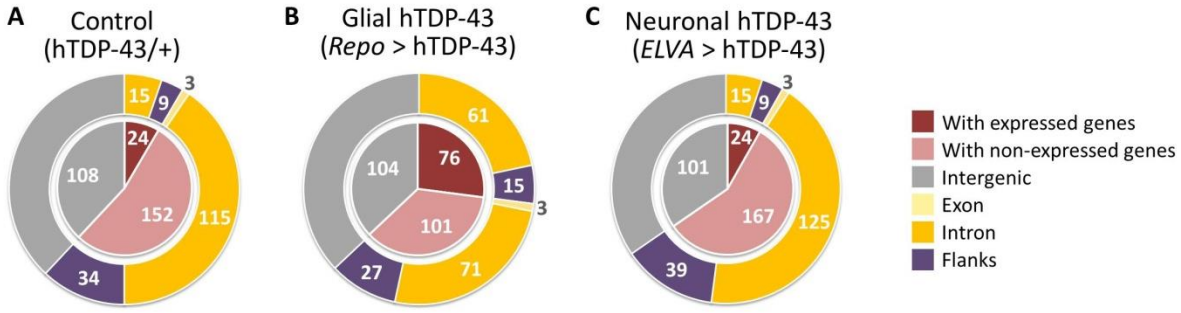



**Figure 5.18 Categorization of *D. melanogaster* autonomous LTR-TE expression candidates by location**

TEs were initially grouped into three categories, with expressed genes, with non-expressed genes, and intergenic. Those genic expression candidates were further binned by location relative to genes, which includes exon, intron, and flanking regions. The digits in each slice or segment denote the number of TE loci in each category.

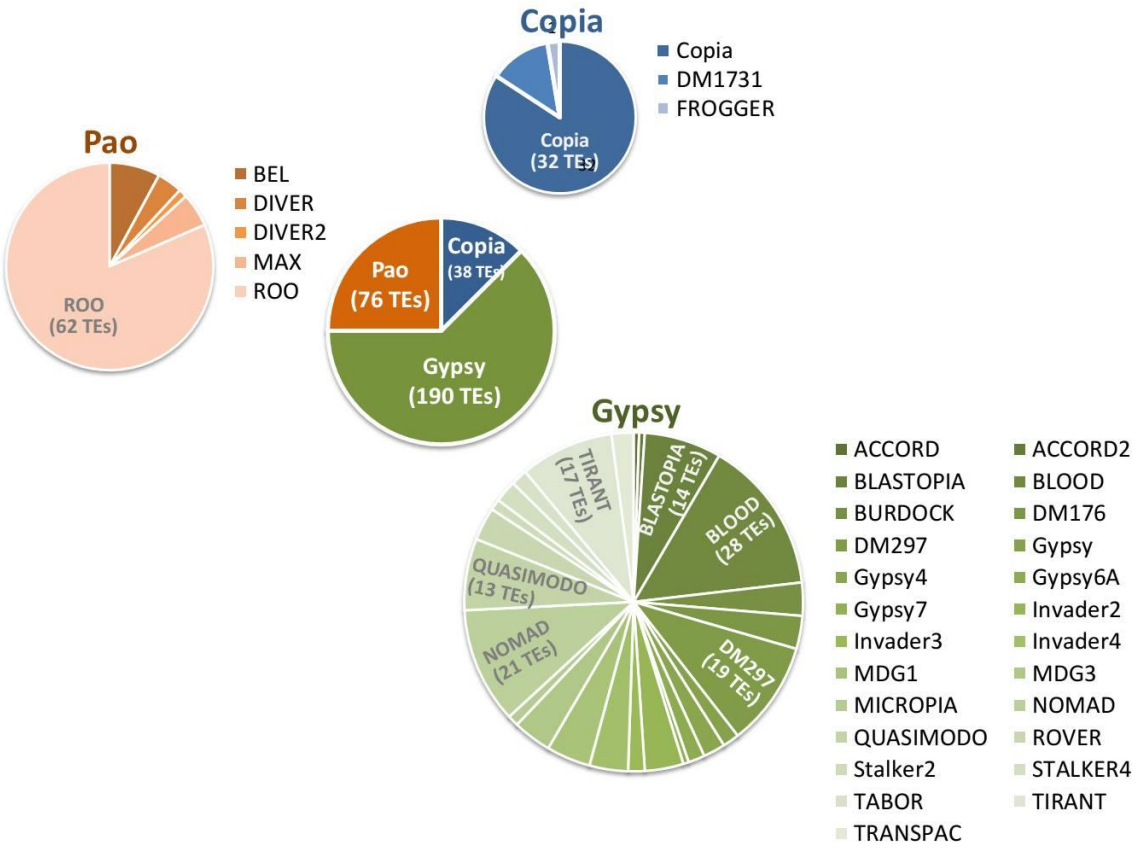



**Figure 5.19 Categorization of the potential autonomous LTR-TE expression candidates in *D. melanogaster* by family**

The total 304 loci pulled from three genotypes were grouped by family (centre), following with further categorization by family shown in the satellite pie graph plotted in proportion to the number of candidate loci. Family obtaining more than ten autonomous candidates were indicated on the corresponding slices.

In *D. melanogaster*, there are 43 LINE families that correspond to 24,133 annotated LINE loci, in which only 344 loci are full-length and potentially competent for autonomous mobilization (Table 5.12). Of the 344 annotated potentially autonomous LINE elements, 305, 293, and 286 were found in the expression candidate pool in control, glial, and neuronal ALS models, respectively. Further investigation on the breadth of coverage across these sites revealed 234 to 245 autonomous candidates that were potentially fully transcribed in each genotype. Two hundred and twenty-six of these LINE loci were identified in all three genotypes (Figure 5.20 A). Analysing at a family level, 21 of the 22 families associated with these autonomous candidates were shared by all three genotypes (Figure 5.20 B). Similar to what has been observed from the location distribution of autonomous LTR-TE candidates in *Drosophila*, the fraction of autonomous LINE candidates co-localized with expressed genes to the total pool of the autonomous candidates in the glial model (31.6%; Figure 5.21 B) was considerably higher than that in the wild-type (11.8%; Figure 5.21 A) and neuronal models (11.3%; Figure 5.21 C). The collected pool of autonomous LINE candidates was mostly contributed by family DNAREP1_DM (62 loci), DOC (58 loci), and FW_DM (47 loci, Figure 5.22).

**Table 5.12 Number of selected *D. melanogaster* TEs at each stage in the workflow of collecting potential origins of autonomous Type I LINE transcripts.**

The competent family denotes those retaining intact reverse transcriptase (RT) domain with putative active sites in the canonical sequence.

| TE subsets | Treatments | # Selected TEs | | |
|---|---|---|---|---|
| | | Full-length | Full-length TE of competent family | >90% coverage |
| **Annotated TEs** | | 367 → | 344 | - |
| **Expr. candidates** | Control | 325 → | 305 → | 245 |
| | Glial hTDP-43 | 311 → | 293 → | 234 |
| | Neuronal hTDP-43 | 306 → | 286 → | 239 |

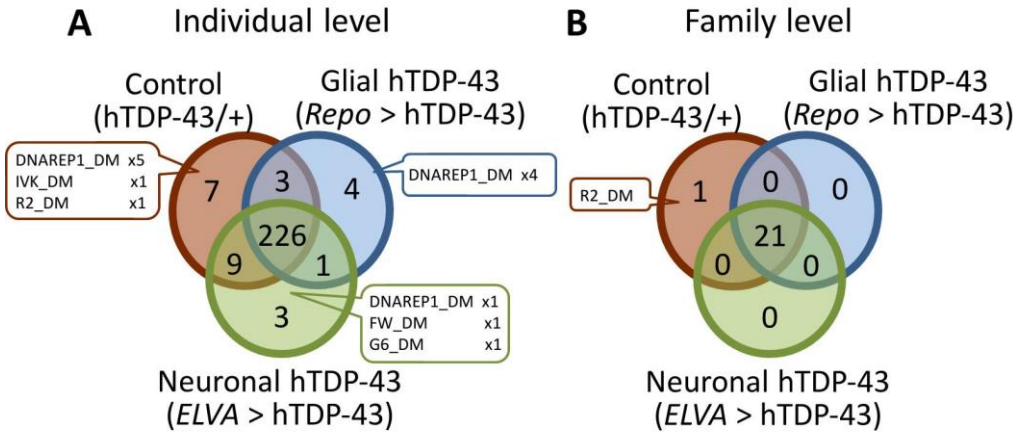



**Figure 5.20 Comparison of potential autonomous LINE expression candidates in *D. melanogaster* across genotypes**

The comparison of the three sets of potential autonomous LINE candidates was illustrated by the Venn diagram at **(A)** individual and **(B)** family levels.

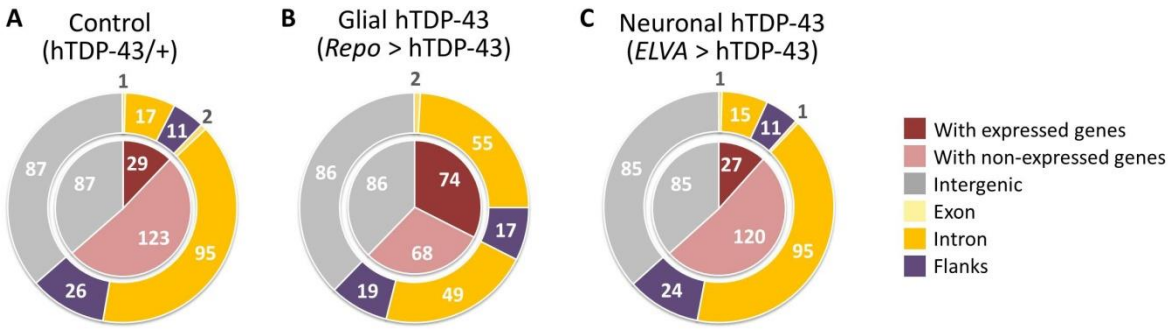



**Figure 5.21 Categorization of *D. melanogaster* autonomous LINE expression candidates by location**

TEs were initially grouped into three categories, with expressed genes, with non-expressed genes, and intergenic. Those genic expression candidates were further binned by location relative to genes, which includes exon, intron, and flanking regions. The digits in each slice or segment denote the number of TE loci in each category.

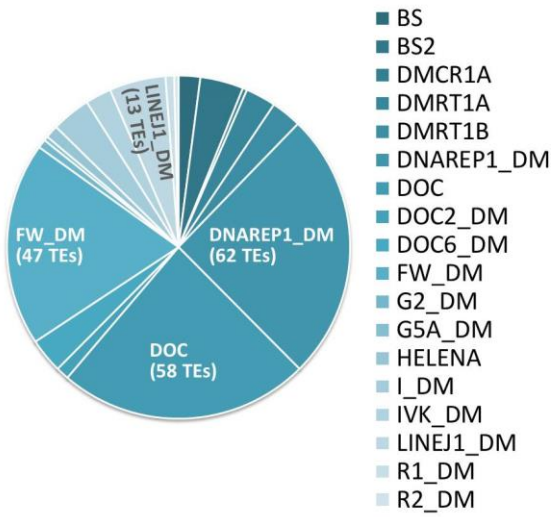



**Figure 5.22 Categorization of the potential autonomous LINE expression candidates in *D. melanogaster* by family**

The total 253 loci pulled from three genotypes were grouped by family. Family obtaining more than ten autonomous candidates were indicated.

## 5.5 Discussion

### 5.5.1 The identification of expression candidates reflects assumptions for *ibm2* and *ddm1*

As a validation for the analysis pipeline established in chapter 3, about 4.5% TE loci in the examined wild-type *A. thaliana* (*Col*) were potentially expressed (Figure 5.1 A). Likewise, 4.3 % of the annotated TEs in the *ibm2 Arabidopsis* were identified as expression candidates (Figure 5.1 B). The identity of wild-type and *ibm2* expression candidates were largely overlapped (Figure 5.4 A), suggesting that permissive regions for TE transcription in the genome were not substantially affected with depletion of the RNA binding protein masking intronic TEs in the gene-TE fused transcripts. On the contrary, the number of expression candidates in *ddm1* is about three times more than wild-type and *ibm2*, while the boundary between expression candidates and non-candidates was blurred by the increased proportion of under-threshold TEs in *ddm1* (Figure 5.1 C). The comparison of the three sets of expression candidates revealed more than 3,000 newly emerged transcribed TE loci, including 58 TE families not transcribed in the wild-type and *ibm2* collection, potentially acquiring transcriptional activity in *ddm1* (Figure 5.4). By use of our analysis pipeline on RNAseq data of wild-type, *ibm2* and *ddm1*, the variance in the number of TE expression candidates and expressed TE families are concordant with the known impact on TE transcriptional activity from loss of functional *IBM2* or *DDM1*. Identification of LTR-TE loci that were likely to produce competent full-length TE transcripts revealed 87 TE loci combined from the three genotypes, but 84 of these loci were uniquely contributed from the *ddm1* dataset (Figure 5.7). This collection includes an intact TE locus derived from *ATCOPIA93* (*EVD*) that has been proven to mobilize in *ddm1* (Tsukahara et al., 2009). These together suggest that the analysis pipeline is capable of identifying transcriptionally active TE loci.

Examination of the location distribution of expression candidates showed that the *ddm1* mutation significantly de-repressed intergenic TEs that would normally remain silenced in the wild-type and *ibm2* mutant backgrounds (Figure 5.6, Figure 5.7). It is notable that 87% to 95% of the *Arabidopsis* expression candidates were trackable loci characterized by unique-mapping reads (Figure 5.7 E), whereas the proportion of trackable expression candidates in grapevine ranged between 69% and 78% (Figure 4.15 E). It is likely that *Arabidopsis* TEs, at least for the Columbia (Col) ecotype, have individually accumulated more sequence divergence than *Vitis vinifera* TEs. Alternatively, it is possible that there are more highly conserved TEs in *V. vinifera* than in the *Arabidopsis* Col ecotype due to higher TE mobilization activity in *V. vinifera*. This might be related to the difference in plant propagation method between these two species: *Arabidopsis* is seed-propagated, whereas wine grapes are clonally propagated. It has been reported that the epigenetic silencing systems are reprogramed in *Arabidopsis* pollen, where the de-repression of TE transcriptional activity in the

vegetative nuclei (VN) results in the accumulation of 21-22 nt siRNAs that induce 'trans-silencing' of the TE loci in the neighbouring sperm nuclei and thus re-enhance the silencing landscape in the sperm nuclei (Slotkin et al., 2009). This finding has been proposed as a mechanism to prevent TE transposition and promote trans-generational TE silencing in seed-propagated plants. By contrast, clonal-propagated plants are more likely to retain somatic cells that have accumulated new TE insertions generation by generation, and hence more highly conserved TE loci in these plants than the seed-propagated plants.

### 5.5.2 Location distribution of *A. thaliana* expression candidates in the wild-type and *ibm2* background reveals location bias towards expressed genes

The location distribution of expression candidates in *Arabidopsis* resembles that in the grapevine. The majority of wild-type and *ibm2* TE expression candidates co-localize with expressed genes (Figure 5.6 C, D). Notably, most of the annotated intragenic TEs were found overlapping with exon in this study, whereas Le et al. (2015) revealed that 85% of *Arabidopsis* intragenic TEs are within an intron. This contradiction is likely due to differences in the gene and TE annotation files between this study and ours. As mentioned in section 5.3, the *A. thaliana* reference genome TAIR10 and gene annotation was obtained from Ensembl Plants (https://plants.ensembl.org/Arabidopsis_thaliana/Info/Index), while the TE GTF file was annotated by Jin et al. (2015) and is available from the lab web page of Prof. Molly Hammell (http://hammelllab.labsites.cshl.edu/). The gene GTF file from Ensembl Plants records 32,833 annotated genes, including 29,352 protein-coding and ncRNA-coding genes, while the TE GTF file contains 31,189 annotated TEs. Le and colleagues (2015) acquired 27,600 genes annotated as 'protein-coding' or 'ncRNA' from The Arabidopsis Information Resource (https://www.arabidopsis.org/) and established their own TE annotation, which contains 7,187 TE loci including fragmented TEs longer than 50 bp, independent of Jin et al. (2015). Since the categorization of intronic and exonic TE in this chapter follows methods in Le et al. (2015), the striking difference in the numbers of annotated TEs could have largely contributed to the discordance of TE distribution in the *Arabidopsis* genome. Besides, based on the annotation by Jin et al. (2015), considerable numbers of long TEs overlap with multiple exons and introns and, therefore, were preferentially categorized as exonic TEs. Nevertheless, this discrepancy doesn't change the conclusion that wild-type and *ibm2* TE expression candidates were predominantly associated with expressed genes, and in light of our data from the grapevine, this is suggestive that this location bias might be common across plant genomes.

### 5.5.3 The scale of potentially transcribed and un-trackable TE loci implies constant TE activation in *D. melanogaster*

The RNAseq data of the healthy and neuronal degenerated *D. melanogaster* appears to reveal a scenario of TE transcriptional activity that is markedly different from *Arabidopsis* and grapevine. About half of the annotated TEs were identified as expression candidates in healthy flies, and another 31% of the annotated loci were found to be transcribed under our threshold limits, leaving only 18% of the total annotated loci that were completely silenced (Figure 5.11 A). This suggests a high proportion of *Drosophila* TEs are expressed at a basal level, and there is a blurry boundary between transcriptionally active and inactive TEs. Compared with the healthy model, *Drosophila* with glial and neuronal hTDP-43 ectopic expression revealed a similar proportion of expression candidates (45% - 46%) and under-threshold TEs (31% - 34%) relative to total annotated TEs (Figure 5.11 B, C). Unlike the situation in *Arabidopsis* and grapevine, the *Drosophila* expression candidates were mostly un-trackable (Table 5.7, Figure 5.16 E), implying that these TEs are highly conserved, possibly due to very recent or constant mobilization of these TEs in *Drosophila*, and therefore the analysis pipeline was not able to distinguish them to the level of individual loci. TEs in wild-type *D. melanogaster* were estimated to contribute 0.57 insertions and 0.037 deletions per generation (Mérel et al., 2020). This transposition rate is comparable with *ATCOPIA93*'s (*EVD*) 0.66 transposition rate per generation predicted from a population of *A. thaliana* epigenetic recombinant inbred lines (epiRILs; Quadrana et al., 2019), suggesting that the relatively frequent TE transposition in *D. melanogaster* might have contributed to the high identity across individual TE loci. Even if the un-trackable expression candidates were ignored, about 30% of the annotated TEs were aligned with unique-mapping reads (Table 5.7). Compared with the absolute numbers of grapevines and *Arabidopsis* trackable expression candidates and with their genome size (*V. vinifera*: ~500Mb; *A. thaliana*: 135 Mb; *D. melanogaster*: 123Mb), it seems that *D. melanogaster* constantly bears the risk of TE mobilization in the neuronal cells.

### 5.5.4 Identification of *D. melanogaster* autonomous expression candidates might facilitate the study of ALS pathogenesis

Such numbers of transcriptionally active TE loci might explain why the ectopic expression of hTDP-43 in glial and neuronal cells didn't substantially increase the number of expression candidates and why the numbers of newly emerged expression candidates in the neuronal degenerated flies were relatively low comparing to total expression candidate pool and in comparison with our findings in the *Arabidopsis ddm1* mutant. In addition, although different transcriptional levels might exist in different genotypes, the potential origins of autonomous LTR-TE and LINE transcripts were mostly conserved across the three genotypes (Figure 5.17, Figure 5.20). Since the overall accumulation of

Gypsy transcripts was enhanced in the glial hTDP-43 flies (Krug et al., 2017), it is likely that the ectopic hTDP-3 expression promoted Gypsy transcript accumulation that was mainly derived from existing transcriptionally active Gypsy loci in healthy flies.

Although Krug et al. (2017) showed that Gypsy-specific siRNAs that were introduced into the glial model significantly improved the pathological symptoms of ALS, it remained unclear that which genes were under the impact of the increased transcriptional activity of Gypsy. Identification of genes co-localized with expressed Gypsy loci might provide clues of affected genes, which can shed light on the mechanism of ALS pathogenesis. However, the individual Gypsy loci that contributed to the elevated transcriptional level were not able to be located in the genome based on the analysis approach in Krug et al. (2017). In contrast, the location distribution of expressed TE loci can be acquired by use of our analysis pipeline. The annotated TEs of *D. melanogaster* are almost equally found in intergenic and genic regions, with genic TEs predominately found within introns (Figure 5.15 A). Unlike the case we have observed in grapevine and *Arabidopsis*, *Drosophila* expression candidates do not appear to exhibit a noticeable numerical location bias towards genic regions, despite statistical tests showing significance (Figure 5.15, Figure 5.16). While most of the genic expression candidates of *Arabidopsis* and grapevine were co-localized with expressed genes, a large proportion of fruit fly expression candidates co-localized with inactive genes. Without further investigation of the location bias by TE family, it is unclear whether location bias of expression candidates is widely absent in most of the TE families in *D. melanogaster* or whether any bias is likely to be specific to individual TE families whose location bias has been averaged out. Nonetheless, the proportion of expression candidates within introns of expressed genes was substantially elevated in the glial-hTDP-43 model (Figure 5.15 C, Figure 5.16 C), which has been correlated with increased severity of disability and mortality (Krug et al. 2017). Likewise, the LTR-TE and LINE loci potentially producing competent transcripts were more frequently found in the glial ALS model (Figure 5.18, Figure 5.21). Besides Gypsy elements, other TE families might also play important roles in causing the disease (Figure 5.19, Figure 5.22). Therefore, the interactions between these active intronic TEs and the expressed host genes might help shed light on ALS pathogenesis.

## 5.6 Conclusions

This chapter demonstrates that the analysis pipeline built in chapter 2 and chapter 3 can be applied to multiple species. In *Arabidopsis* and *Drosophila* RNAseq data, the scale of potentially expressed TE loci was narrowed to 4% and 50% of all annotated TEs, respectively. The efficiency of narrowing the candidate pool is likely to be species-specific depending on the transposition rates. The sensitivity of our approach, in terms of identification of trackable expression candidates, is compromised by increasing numbers of recently mobile elements that tend to be identified as un-trackable elements, the abundance of which varies across species and kingdoms. Our analysis approach highlighted that the location bias of *Arabidopsis* expression candidates is similar to what have observed in *V. vinifera* and successfully reflected the expectation for plants that are compromised in the epigenetic mechanisms targeting and silencing TE loci in the *ddm1* mutant, where remarkably increased number of expression candidates in the intragenic region were demonstrated. Furthermore, with the identification of transcriptionally active TE loci and the co-localized genes, the utilization of our analysis pipeline might facilitate the interrogation of genes epigenetically affected by increased expression activity of co-localized Gypsy elements in the *Drosophila* model of ALS, hence a better understanding of ALS pathogenesis.

While our analysis pipeline (based on short-read sequencing data) improved the granularity in the interpretation of TE transcriptional activity to the level of individual TE loci, the inadequacy of short-read sequencing in resolving alignment of TE-rich regions leaves doubts about the liability of the putative complete transcription derived from structurally autonomous LTR-TE loci. It is also unclear whether TEs co-localized with genes may influence alternative splicing behaviour and thus further influence gene expression dynamics. To address these, in the next chapter, we deployed Oxford Nanopore Technology long-read sequence that enables sequencing of full-length transcripts as a single read to capture intact transcripts and validate the findings obtained from short-read Illumina type sequencing approaches.

# Chapter 6

# Analysis of TE transcriptional activity using long-read cDNA sequencing

## 6.1 Overview

The maturating of long-read sequencing technology provides a powerful solution to resolve the genomic and transcriptomic complexity contributed by TEs. Here the ability of ONT cDNA sequencing in a TE-oriented study was examined on grapevine embryogenic callus treated with the Mock/wound-like procedure and live yeast (*Hanseniaspora uvarum*). Based on the same experimental design, the ONT cDNA library was comparable with the Illumina Truseq RNA sequencing (RNAseq) libraries reported in chapter 2 in terms of the total sequencing bases, the alignment coverage, as well as the quantification level of genes and TE families. Combined analysis of the ONT and Illumina datasets to identify expressed TEs resulted in a more confident collection of expressed TE loci in terms of breadth coverage and ONT read counts. The ONT cDNA data proved to be able to validate the observations from the Illumina libraries. Furthermore, with the advantage of acquiring intact sequence information of a transcript, this new technology facilitates the systematic study of alternative splicing directly associated with TEs. Intron retention (IR) overlapping with annotated TEs was particularly well correlated with the exposure of premature termination codons (PTC). It has been reported that heterochromatic hallmarks, which are commonly associated with TE sequences, participate in alternative splicing. Therefore, it is possible that TEs included in the IR might influence alternative splicing through altering epigenetic dynamics and thus expose PTC provided by TEs. Finally, the ONT cDNA data was capable of capturing full-length transcripts spanning through autonomous TEs loci derived from Gypsy-V1 and hAT-7. Nevertheless, the level of the competent TE transcripts was extremely low, and the de novo full transcription of the most active TEs (Copia-3 and Copia-23) seen from the Illumina data were not detected in the ONT dataset. A thorough survey for stress-responsive cis-regulatory elements (CREs) on the long-terminal repeats (LTRs) of Copia-3, Copia-23, and Gypsy-V1 suggested that the applied stress treatments may not be sufficient to boost de novo transcription of autonomous TEs alone. Overall, it appears that choosing proper stimuli according to the annotated CREs coupled with inhibition of epigenetic silencing may be required to efficiently activate transcriptional activation of autonomous TEs.

## 6.2 Introduction

The repetitive and self-proliferating nature of TEs has long been problematic for deciphering genomic or transcriptomic sequencing results. In the field of Next Generation Sequencing (NGS), the most common sequencing technologies produce sequencing reads shorter than 300bp, while other platforms such as 454 (Roche) generate reads no longer than 1kb (van Dijk et al., 2014). Given that an autonomous TE ordinarily is longer than the sequence limitations of short-read second-generation sequencing technology, reads derived from TEs are often unable to be mapped to a single location in the genome. This is exacerbated by the genomic complexity associated with the nesting of repeat sequences that can often span multiple megabases of genomic space (Jedlicka et al., 2019). The development of third-generation sequencing technologies has opened up opportunities to effectively probe the complexity of the 'dark matter' in genomes. The Pacific Biosciences (PacBio) 3$^{rd}$ generation sequencing platform can generate single read-lengths up to 150 kb, while Oxford Nanopore Technologies (ONT) can reach >4Mb (Kilburn et al., 2020). With the read-length and sequence volume capacity of long-read sequencing, there is a chance to resolve the ambiguity in repeat-rich regions. So far, there have been several examples of TE-oriented studies using ultra long-read genomic sequencing (Shahid and Slotkin, 2020), yet the application in detecting TE transcripts is, to date, rare. Therefore, this chapter aims at exploring the ability of ONT cDNA sequencing in detecting active TEs and validating our observations of the Illumina dataset presented in earlier chapters.

### 6.2.1 Oxford Nanopore cDNA Sequencing

An ONT nanopore is a current-carrying nano-scale pore that allows individual DNA or RNA molecules to pass, and in doing so, interrupt ionic flow through the pore. The current change is measured and interrogated to determine the sequence of bases passing through each pore. To detect gene expression, researchers can use cDNA-based protocols which convert mRNA into cDNA during library preparation. Alternatively, RNA molecules can be sequenced directly based on the protocol of direct-RNA sequencing. Direct RNA sequencing has the added advantage of detection of RNA modification at single-nucleotide resolution (Lorenz et al., 2020; Parker et al., 2020), while ONT cDNA sequencing provides sequencing of full-length cDNA transcripts at high yields (https://store.nanoporetech.com/cdna-and-direct-rna/). With the enrichment of full-length cDNA, the PCR-based cDNA sequencing approach particularly benefits TE-oriented study, in which the level of full-length transcripts can be determined, hence a potentially valuable approach to identify autonomous TE transcripts that may be produced at extremely low levels relative to genes.

Using the Oxford Nanopore PCR-cDNA approach, mRNA can be captured and reverse transcribed with the presence of VN primer (VNP) consisting of a string of 10 deoxythymidine (dT) following with

two VN nucleotides, where V denotes A, C, or G and N represents A, T, C, or G. The reverse transcription would be completed with the aid of strand-switching primer (SSP). The VNP and SSP serve as docks for PCR primer annealing. With the DNA polymerase processivity of 50 seconds per Kb, the time period of the extension step in the PCR cycle is decisive in determining the overall length of enriched cDNA molecules. An exonuclease I digestion following the PCR step removes single-stranded primers and linear single-stranded DNA in the sample, leaving a double-stranded product of full-length cDNA. This measure was found to effectively reduce the blocking of the nanopores and resulting in a boost in sequencing throughput (Nanoporetech Community). Purified PCR product would then attached to sequencing adapters containing motor proteins that activate nanopores upon attachment, thus triggering the flow of DNA molecule through the pore.

### 6.2.2 Key tools in the processing and analysis workflow

#### 6.2.2.1 Basecalling using Guppy

The raw electrical signal of nanopore sequencing resembles squiggles, which can be translated into nucleotide sequences by sophisticated basecalling tools. The computational strategy of basecalling is crucial to the interpretation of the current changes. Although ONT sequencing is originally error-prone (10-15%) compared to short-read sequencing (Shahid and Slotkin, 2020), its sequencing accuracy is continually improving through improvements in nanopore chemistry, library preparation and basecalling algorithms (Jain et al., 2017; Rang et al., 2018; Volden et al., 2018). The computational strategy for base calling has evolved from using the statistics (e.g. the mean, standard deviation, and duration) of segments of the electrical signal to direct use of the raw signal with Recurrent Neural Network (RNN)-based algorithms (Rang et al., 2018). The ONT-developed basecaller, Albacore, has been adopting the latest computational strategies to improve basecalling accuracy. Inheriting the computational advantages from CPU-based Albacore, ONT's Guppy basecaller outperforms in terms of computing speed by using GPUs (Wick et al., 2019).

#### 6.2.2.2 Selection of full-length reads using Pychopper2

There are a huge variety of analysis tools developed for ONT to enrich the full-length sequencing reads. Pychopper2 uses the aforementioned VNP and SSP primers as tokens for the selection of full-length cDNA reads before performing adapter trimming (Oxford Nanopore Technologies, 2020a). Pychopper2 first looks for the presence of VNP or SSP sequences in each sequencing read. Sequencing of a full-length cDNA would involve the leading strand of full-length cDNA, which is sequenced in the order of SSP, cDNA sequence, and then VNP. In cases of two cDNA reads being sequenced one after the other, a fused sequencing read would be formed. In addition, there would be more than two flanking primers detected by Pychopper2 in the fused read. Pychopper2 is then able to split the read into segments by recognizing the boundary of two consecutive primer hits. As a

result, Pychopper2 not only identifies and trims full-length ONT cDNA reads but also rescues fused reads to maximize the size of qualified read pools.

#### 6.2.2.3 Alignment of reads using minimap2

The development of aligners for ONT sequencing data is an active field. As new software emerges, several tools originally designed for short-read sequencing have also been updated with parameters optimized for long-read mapping. For ONT cDNA or direct RNA sequencing data, the awareness of splicing is of critical importance for alignment accuracy. Published in 2008, GMAP is a splicing-aware aligner developed for mapping mRNA and expressed sequence tags (ESTs) in the form of conventional cDNA sequencing data (Wu and Watanabe, 2005). It was further applied in analysing the benchmarking of ONT direct RNA sequencing (Garalde et al., 2018). Nonetheless, it was reported that only 68.7% to 84.1% of the annotated splice junctions were correctly aligned by GMAP when evaluated using human ONT Direct RNA sequencing data (Li, 2018). On the contrary, using the same dataset, the software minimap2, which was initially designed for genomic DNA alignment, was able to correctly predict 94.2% of the annotated splice junctions by invoking minor modifications in the base algorithm (Li, 2018). Furthermore, minimap2 was found to be 160 times faster than GMAP when working on the same data. These have made minimap2 one of the most commonly used aligners for ONT cDNA or RNA sequencing data.

#### 6.2.2.4 Detection of alternative splicing using FLAIR pipeline

ONT long-read RNA or cDNA sequencing facilitates studies of alternative splicing because each RNA or cDNA sequencing read of this technology contains intact information of an RNA transcript. However, most software has been designed for working with second-generation short-read sequencing technologies having high base accuracies and are generally not suitable for a platform such as ONT 1D cDNA sequencing that has high base inaccuracies (Weirather et al., 2017). Using such software with ‘noisy’ ONT-derived sequence reads often results in the exclusion of important data from isoform analysis (Tang et al., 2020). Recently, several tools have been developed specifically for ONT data. The key steps of these tools include the alignment of multiple reads to the reference genome, correction of splice sites, grouping isoforms by splice junctions, and establishment of consensus isoform datasets. Among these steps, defining the correct and consensus splice sites, particularly for unannotated splice sites, is of critical importance.

The sequencing errors of the ONT platform are usually reflected as INDELs or small gaps in the alignment. Different to the conventional Multiple Sequence Alignment (MSA) that tends to linearize multiple alignment scenarios by introducing gaps, Partial Order Alignment (POA) graph approaches takes all inconsistencies into consideration by generating conjunction nodes and forks in a graph-like order without degeneracy of the data information (Lee et al., 2002). To improve the identification

accuracy of the consensus isoforms from ONT data, the Mandalorion pipeline (Byrne et al., 2017) invokes the POA algorithm by implementing the package Racon (Vaser et al., 2017) in the pipeline. Inspired by Mandalorion, the ONT-developed pipeline Pinfish uses Racon to polish consensus isoforms predefined by the median exon boundaries from all reads in the isoform clusters (Oxford Nanopore Technologies, 2020b). In a recent comparison of Mandalorion and Pinfish with the FLAIR pipeline, FLAIR was found to outperform both in ONT-based isoform identification in terms of sensitivity and precision (Tang et al., 2020). Instead of using Racon, FLAIR fills small gaps to address the INDELs and corrects splice junctions based on either annotated exon boundaries or short-read data. In addition, FLAIR has built-in tools to quantify isoform expression, analyse types of alternative splicing (AS), predict isoform productivity, and perform a statistical test of differential usage. The combined use of the tools in FLAIR facilitates the systematic study of isoform usage with the estimation of the biological outcome.

### 6.2.3 Association between TEs and alternative splicing

Alternative splicing can attribute transcriptome and proteome diversity across tissue types, species, or environmental conditions. Alternative Splicing includes, but is not limited to, alternative 5' splice site (Alt5), alternative 3' splice site (Alt3), intron retention (IR), and exon skipping (ES). With respect to gene-related isoforms, TE-associated alternative splicing is of great interest in this study. As previously reported in chapter 3 and chapter 4, an impressive proportion of TE expression candidates sit within the intron of expressed genes (Figure 4.12), and over 30% of genes contain TEs in introns (Figure 4.1). Generally speaking, these genes are less likely to be highly expressed (Figure 4.2). Furthermore, intragenic TEs may negatively influence gene expression by participating in alternative splicing.

Autonomous LINE and non-autonomous SINE retrotransposons, such as L1 and Alu, have been considered responsible for several aberrant alternative splicing linked to human disease (Ayarpadikannan et al., 2015). Two Alu elements that were artificially added into an intron in opposite orientation resulted in secondary structure formation of the transcripts and exon skipping (Lev-Maor et al., 2008). By examining the insertions polymorphism of endogenous retrotransposons (ERV) across multiple mouse strains, Li et al. (2012) found that the presence of an ERV in an intron was responsible for the increases in premature transcriptional termination. The truncated transcripts stopped at 1.5Kb upstream of the intronic ERV insertion. The premature stop signal was conserved across mouse strains, yet only the strain with intronic ERV demonstrating 49-fold increases of the truncated isoform. This indicates the intronic TE affected the use of alternative polyadenylation signals, albeit the signal was not within the TE. Insertion polymorphisms of long retrotransposons within introns of maize waxy (Wx) gene resulted in three waxy alleles expressed at low levels

comparing to the wild-type allele (Varagona et al., 1992). These insertions increased the intron size by 40 to 60 fold. Although the TE-containing introns were spliced from Wx pre-mRNA, these elements disrupted recognition of the constitutive splice sites and caused the exclusion of flanked exons from mRNA. It seems that, in addition to directly providing cryptic alternative splice sites or premature termination codons, intragenic TEs may participate in splicing regulation with other mechanisms.

Since the splicing of pre-mRNA is executed co-transcriptionally, it has been proposed that DNA methylation and chromatin modification involves in alternative splicing (Lev Maor et al., 2015; Saint-André et al., 2011). In oil palm, loss of epigenetic silencing on an intronic LINE was found underlying the abnormal exon-skipping, which introduces premature termination stop site of a homeotic gene and leads to the fertile mantled phenotype (Ong-Abdullah et al., 2015). Saint-André and colleagues (Saint-André et al., 2011) found that the enrichment of the heterochromatic hallmark H3K9me3, which usually implicates the presence of TEs underneath, strongly associated with exon skipping.

Overall, these studies highlight the association between TEs and alternative splicing, in which intragenic TEs may participate in alternative splicing by providing cryptic splicing signals or through attracting epigenetic modification. This suggests that besides functioning as CREs or alternative promoters (see chapter 3.2), TEs may also regulate gene expression through alternative splicing. Therefore, in addition to using ONT to validate the findings in Illumina data, this chapter also investigates the linkage between TEs and alternative splicing in grapevines. Since no epigenetic analysis was included in this chapter, the alternative splicing analysis focuses on the survey of alternative splicing features overlapping with TEs, as well as on the prediction of isoform productivity with TE-related alternative splicing.

## 6.3 Methods

### 6.3.1 Stress treatment

The embryogenic callus was established as described in chapter 3. The embryogenic callus was subjected to the mock treatment that would confer a wound type response and live *H. uvarum* (denoted as yeast) treatment following the steps documented in chapter 2. The callus of mock treatment was harvested after 12 hours of recovering on the $C_1^P$ plate, whereas the yeast treatment involves 12 hours of continuous incubation with the yeast before harvesting.

### 6.3.2 RNA extraction, ONT cDNA library preparation, and sequencing

Total RNA was extracted and separated using the NORGEN Plant microRNA purification kit (Norgen Biotek) according to the manufacturer's instruction. Genomic DNA contamination was removed with

the standard protocol of the TURBO DNA-free kit (Thermo Fisher). Amplification of the grapevine *ACTIN* gene with 35 PCR cycles following with electrophoresis was conducted to confirm that there's no detectable genomic DNA in the RNA samples. The RNA quantity was measured by Qubit RNA BR (Broad-Range) Assay Kit (Thermo Fisher), and the quality was examined using Agilent 2100 Bioanalyzer, in which the resulting RIN value of each library was above 8.

The cDNA library was prepared following the protocol of the Oxford Nanopore cDNA-PCR kit (SQK-PCS109). Briefly, 50 ng of total RNA was reverse transcribed using Maxima H Minus Reverse Transcriptase (Thermo Fisher) with the presence of VN primer (VNP). In the same reaction tube, the reverse transcription was completed with a strand-switching step by using the strand-switching primer (SSP) and the strand-switching activity of the reverse transcriptase. The resulting full-length cDNA was further enriched by PCR, which involved 12-13 amplification cycles, each with 6 min of extension step. The amplified cDNA was purified by AMPure XP beads before ligation of the 1D sequencing adapters. Finally, the cDNA library was loaded onto an R9.4.1 MinION flow cell and then sequenced using MinKNOW (version 18.12) control software for raw data collection only. Basecalling was carried out offline using Guppy (version 3.2.1) as described below.

### 6.3.3 Processing of sequencing data

Raw fast5 reads were basecalled using Guppy (https://community.nanoporetech.com). The resulting fastq files were processed by Pychopper2 (Oxford Nanopore Technologies, 2020a) to capture full-length reads and remove adapter sequences. The full-length reads were mapped to the 12X PN40024 *V. vinifera* reference genome using minimap2 (Li, 2018) with the pre-set option `-ax splice` for long-read splice alignment. This default setting output the alignment with the best mapping score as primary alignment and at most five secondary alignments that could align reasonably well. In cases of multiple equally good alignments, minimap2 would randomly designate one as primary alignment and the rest as secondary alignment. Therefore this setting is suitable for general gene analysis. For mapping self-proliferating and highly repetitive TE sequences, the output of up to 100 secondary alignments was allowed for individual TE analysis by using `-N 100 -ax splice`. For analysis based on TE family level, the ONT reads were mapped to the set of 232 canonical TE sequences by running default `minimap2 -ax splice`, before using `bedtools coverage` to quantify mapped reads at the family level. For genes, based on grapevine gene annotation (version 2.1) downloaded from the Grape Genome Database at CRIBI (http://genomics.cribi.unipd.it/grape/), FLAIR pipeline was then applied to obtain high fidelity isoforms by running `flair correct` and `flair collapse`, where quantification of the isoform expression level was conducted by calling `flair quantify` and reported as transcripts per million (TPM). The TPM of isoforms derived

from the same gene was then summed up at the gene level for the overall quantification of gene expression. For individual TEs, ONT reads overlapping with TEs in sense orientation were collected and quantified by `bedtools intersect` and `bedtools coverage` (Quinlan and Hall, 2010) as described in chapter 3.

At each processing stage, the mapping statistics, including mapped read count, mapped bases, and N50 were generated by NanoPlot (De Coster et al., 2018), except that the sense-oriented reads and the read bases overlapping with individual TEs was calculated as per the pipeline in chapter 2 while the corresponding N50 was estimated by NanoStat (De Coster et al., 2018).

### 6.3.4 Comparison between ONT and Illumina data

To compare the correlation between ONT and Illumina platforms at the individual gene level, genes' TPM (logarithmically transformed) from ONT was plotted against FPKM (logarithmically transformed) from Illumina data, while the correlation was tested using Spearman's correlation coefficient. The same approach was applied to TE families, in which the expression level of each TE family was obtained from TEtranscripts (see chapter 3) for the Illumina dataset and from alignment with the canonical sequences followed by bedtools coverage analysis (see 6.3.3) for ONT data.

### 6.3.5 Identification of expressed TEs and genes

Individual TE loci overlapping with at least one ONT read were collected to perform an intersection with the expression candidates obtained from Illumina libraries at a time-point of 12 hours. The intersected TE loci were supported by both sequencing platforms and thus considered as expressed TEs. All TEs with at least one ONT read were further examined for the proportion of annotated TE feature covered by ONT read (denoted as breadth coverage) and their ONT read count generated by bedtools coverage, in which multi-mapping reads would be counted multiple times. This analysis was then plotted as a scatter plot and a density plot to reveal the different distribution of breadth coverage and read counts between TE loci only collected by ONT data and TE loci supported by both ONT and Illumina datasets.

To identify expressed genes, an intersection between annotated genes with ONT TPM above 1 and genes with Illumina FPKM over 1 was performed. Genes having overlapping data from both platforms were considered transcriptionally active genes.

### 6.3.6 Validation of previous findings in Illumina data

The key tasks of this section were to validate the location bias of expression candidates described in chapter 4 and the negative correlation between TEs and gene expression levels reported in chapter

5. Using the same analysis approach in chapter 4, the former task involves the characterization of expressed TEs by location, integrity, and distinctness. As described in chapters 2 and 3, the distinctness, denoted as trackable or un-trackable, of a TE locus means whether this locus was overlapping with unique-mapping read or not, respectively. Using Illumina data, TE loci having unique-mapping reads can be identified by htseq-count (see chapter 3), which is a tool that can be tuned to take unique-mapping read into consideration by recognizing the `NH` SAM field tag. The number following the `NH` tag would represent the number of distinct genome sites mappable by a read (e.g. A unique-mapping read would be flagged with `NH:1`). Not every aligner (e.g. minimap2) reports this tag. However, since equally good alignments would be randomly designated as primary or secondary alignment by minimap2, and minimap2 doesn't report SAM field tag regarding the alignment uniqueness, the recognition of unique- and multi-mapping reads was performed by 'home-brew' scripts (Appendix D.4) that counted the number of entries in the SAM file for each TE-overlapping read. As to the later validation task, the analysis was based on the scripts used for chapter 4, except that the gene expression level was quantified as TPM as described in 6.3.3.

### 6.3.7 Alternative splicing analysis

The tool `flair diffSplice` embedded in the FLAIR pipeline was used to group alternative splicing into four types, including alternative 5' splicing (Alt5), alternative 3' splicing (Alt3), intron retention (IR) and exon skipping (ES). In addition, the productivity of each isoform was estimated by FLAIR's `predictProductivity`, which predicts four types of productivity, including productive isoform (PRO), presence of premature termination codon (PTC), absence of start codon (NGO) and absence of stop codon (NST). To further capture AS feature directly related to TEs, `bedtools intersect` was utilised to identify alternative splicing features overlapping with annotated TE loci. These features would be denoted as TE-associated or TE-related AS features.

### 6.3.8 Identification of autonomous TEs having full transcription.

The strategy to find autonomous TEs having full transcription in the ONT dataset was as same as that in chapter 3, in which autonomous loci with the potential of full transcription was identified by over 90% breadth of coverage across the internal domain (INT) of LTR-TE, the full feature of LINE, or the open reading frame (ORF) of TIR-TEs. Furthermore, with the consecutive transcript information provided by the ONT platform, the aforementioned potential loci were further examined by the presence of ONT read covering the TE feature or domain necessary for autonomous mobilization. This analysis was visualized by plotting the length of ONT reads mapping to the autonomous TE loci against the number of read-bases overlapping with TEs. This is to select TE loci fully covered by individual reads spanning across the necessary regions instead of those covered by co-contribution of

multiple ONT reads. These reads were also surveyed for their transcriptional start and stop sites relative to mapped TE loci. `bedtools intersect` was firstly used to distinguish ONT reads started or ended internally or externally. For those internal reads, if there were more than 10 clipped bases, these starts or ends would be denoted as clipped in Figure 6.18 – Figure 6.24. The genome browse image of qualified TE loci was generated by a local instance of JBrowse v1.16.8 (Buels et al., 2016).

### 6.3.9 Analysis of stress-related cis-regulatory element

To determine which stress types that might efficiently boost potentially autonomous TE's full-length transcriptional activity, the surveys for stress-related cis-regulatory elements (CREs) on canonical LTR sequences of Copia-3, Copia-23 and Gypsy-V1 were conducted using the web-based tools of Plant Promoter Analysis Navigator (PlantPAN 3.0) at http://PlantPAN.itps.ncku.edu.tw (Chow et al., 2019). Firstly a list of cross-species and stress-related CREs was established using the TF/TFBS Search tool with keywords listed in Appendix C.7. This generated a list of 100 transcriptional factors (TF) binding motifs (Appendix C.7). Transcription factor binding sites of the canonical LTR sequences of Copia-3, Copia-23 and Gypsy-V1 were predicted by the Promoter Analysis tool of PlantPAN 3.0 based on the cross-species database. The resulting lists of TF binding motifs on LTRs were intersected with the stress-related list (Appendix C.7) to annotate stress-responsive CREs on LTRs. The software Unipro UGENE (Okonechnikov et al., 2012) was used to plot the annotated CREs.

All computational scripts used in these analyses can be found in Appendix D.4.

## 6.4 Results

### 6.4.1 Comparison of Illumina and ONT cDNA sequencing in terms of TE and gene expression quantification

The ONT cDNA sequencing generated 11 and 12 million reads in mock and yeast libraries, respectively (Table 6.1). Using Pychopper to select full-length sequencing reads and remove adapters, about 8.5 and 9.5 million reads were retained in the two libraries, respectively. Over 80% of full-length reads, roughly equivalent to 4.7 to 5.6 billion bases, mapped to *V. vinifera* reference genome. The N50 of the raw sequencing output was 983 and 1013, respectively, in mock and yeast libraries and dropped by ~100bp as the selection of full-length reads and alignment of the sequencing reads proceeded. Among ONT reads mapping to the reference genome, about 1% of these reads overlapped with annotated TE loci, including 23.1 and 27.8 million mapped bases in mock and yeast libraries, respectively. The N50 of TE-mapped reads was above 1,200bp and was on average higher than that of total mapped reads.

Comparison of the processed ONT data yield with that from the Illumina run highlighted that the ONT sequencing run produced 8,571 million bases and 9,917 million bases compared to the 5,200 and 5,700 million bases of Illumina read data used in the analyses in chapters 2 (Figure 6.1 A, D). The ONT advantage in almost doubling sequencing output was not maintained after alignment, but the levels of total genome-mapped bases (Figure 6.1 B, E) and TE-mapped bases (Figure 6.1 C, F) in mock and yeast ONT libraries were still comparable with that in the Illumina libraries.

This ONT dataset was generated from experiments independent of those for Illumina Truseq sequencing described in chapter 3 to 5. In order to understand how similar the two datasets are, gene FPKM and TPM values, respectively, obtained from Illumina and ONT were compared for each gene. For mock treatment, the comparisons all show Spearman's correlation coefficient (ρ) = 0.847 (Figure 6.2 A-C). Similarly, for the yeast experiment, the correlation coefficients (ρ) between the libraries generated by ONT and Illumina platforms are all above 0.8 (Figure 6.2 D-F). When it comes to TE families, Spearman's correlation coefficient (ρ) for TE (Figure 6.3) is 0.64 for mock and 0.58 for yeast experiments. This level of correlation coefficient was interpreted as a moderate correlation in medical research (Mukaka, 2012).

**Table 6.1 Mapping statistics of oxford nanopore (ONT) cDNA sequencing (SQK-109)**

| | Sequenced | | | Adapter removal | | | Total mapped | | | TE-mapped | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Mock** | # Reads | 11,045,240 | 100.00% | # Reads | 8,569,467 | 77.59% | # Reads | 6,876,783 | 62.26% | # Reads | 118,803 | 1.08% |
| | Total bases | 8,571,109,556 | 100.00% | Total bases | 5,124,141,507 | 59.78% | Total bases | 4,725,493,852 | 55.13% | Total bases | 23,178,670 | 0.27% |
| | N50 | 983 | | N50 | 859 | | N50 | 880 | | N50 | 1,221 | |
| **Yeast** | # Reads | 12,324,745 | 100.00% | # Reads | 9,489,301 | 76.99% | # Reads | 7,887,361 | 64.00% | # Reads | 139,564 | 1.13% |
| | Total bases | 9,917,204,153 | 100.00% | Total bases | 5,983,361,974 | 60.33% | Total bases | 5,597,447,311 | 56.44% | Total bases | 27,818,485 | 0.28% |
| | N50 | 1,013 | | N50 | 893 | | N50 | 912 | | N50 | 1,284 | |

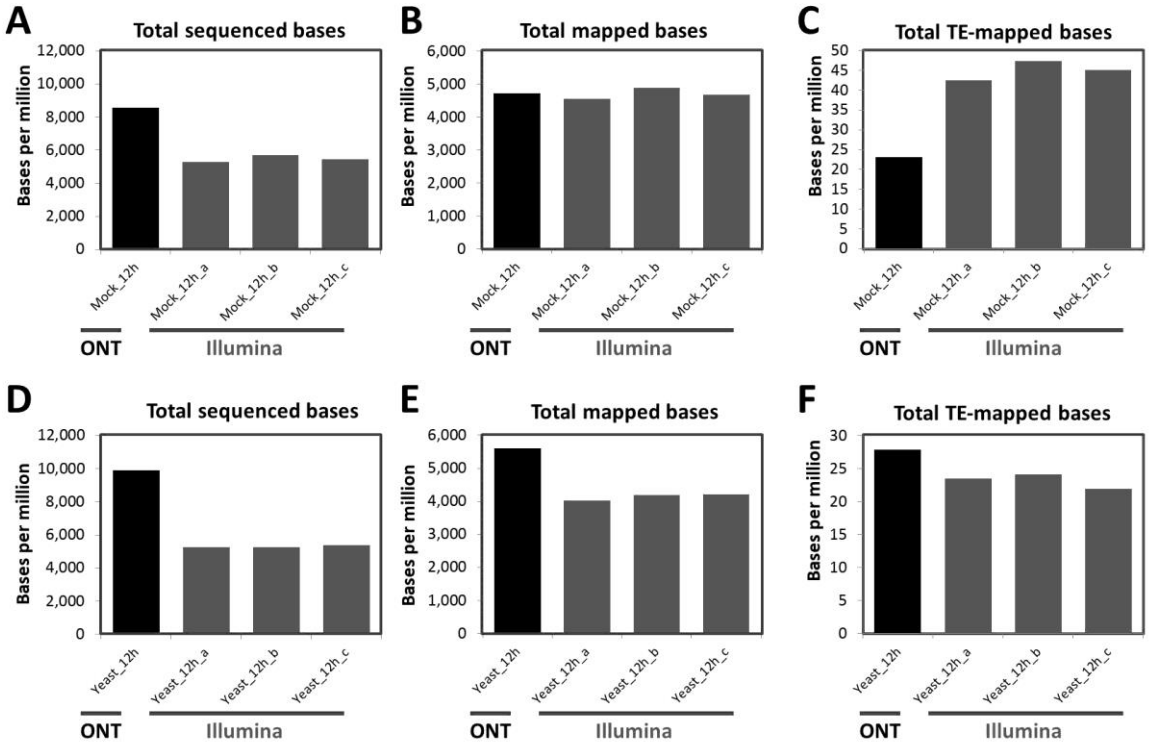



**Figure 6.1 Comparisons of sequencing and alignment output between ONT and Illumina Truseq sequencing libraries**

**(A)-(C)** For mock (12 hours) treatment, the total sequenced bases (A), total mapped bases to the reference genome (B), and total read bases overlapping with TEs (C) were illustrated. **(D)-(F)** For yeast (12 hours) treatment, the comparisons of total sequenced bases, total mapped bases to the reference genome, and total read bases mapping to TEs were demonstrated in (D), (E) and (F), respectively. Of the four libraries shown at the x-axis of each graph, one was sequenced using ONT technology, and the other three were sequenced by Illumina HiSeq2500.

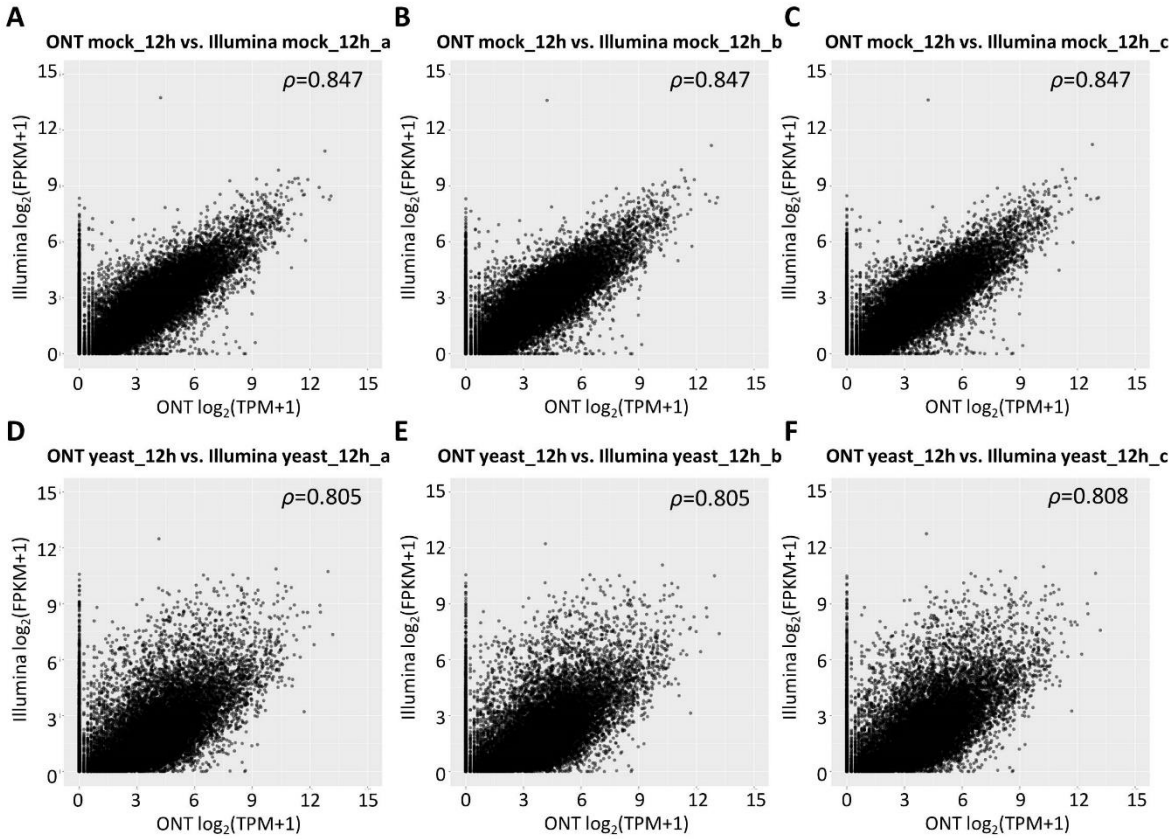



**Figure 6.2 Comparisons between gene expression quantified from ONT and Illumina Truseq sequencing libraries**

**(A)-(C)** For mock (12 hours) treatment, gene expression level quantified from the ONT library was compared to each of the replicates sequenced by Illumina RNAseq. **(D)-(F)** For yeast (12 hours) treatment, the ONT library was compared to each of the libraries sequenced by Illumina RNAseq. The gene expression levels were given as transcripts per million mapped reads (TPM) for the ONT libraries (x-axes) and as fragments per kilobase per million mapped reads (FPKM) for the Illumina libraries (y-axes). Spearman's correlation coefficient ρ was given per comparison, with each point representing gene expression levels.

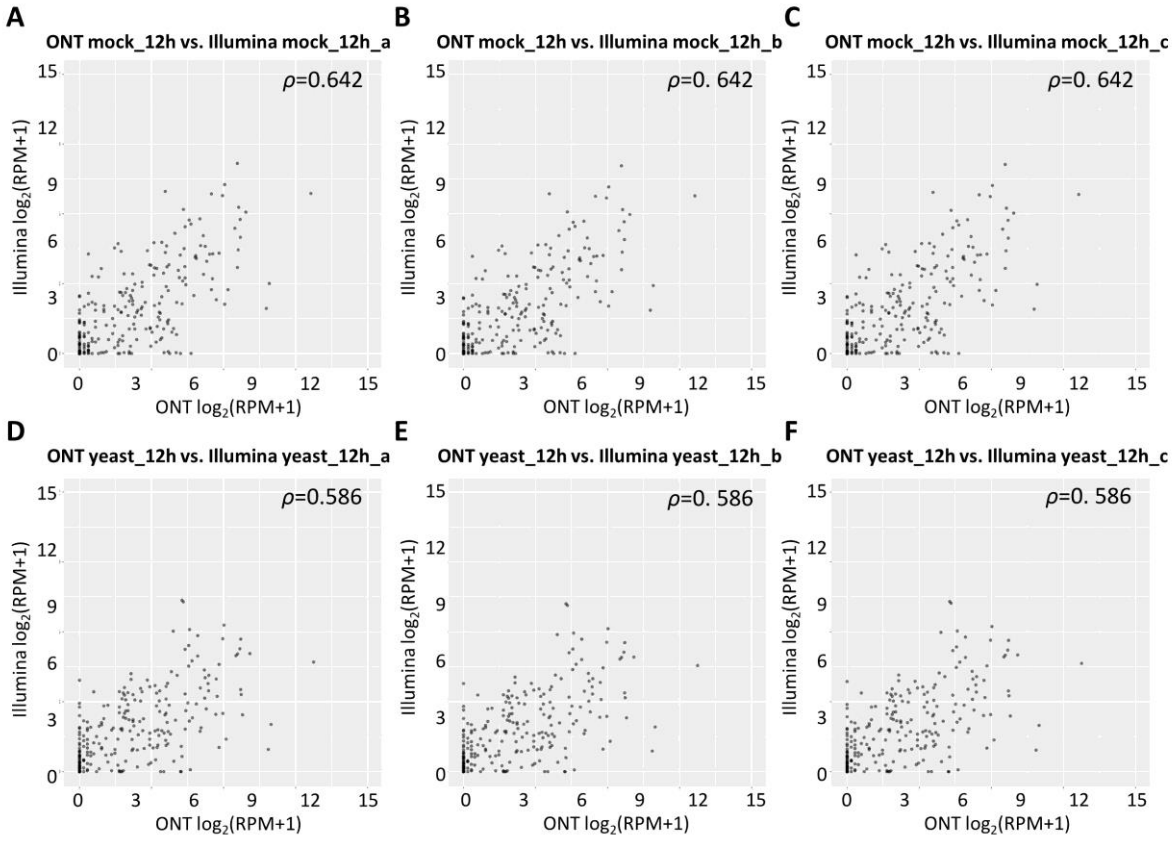



**Figure 6.3 Comparisons between TE family expression quantified from ONT and Illumina Truseq sequencing libraries**

**(A)-(C)** For mock (12 hours) treatment, the expression level of each TE family quantified from the ONT library was compared to each of the replicates sequenced by Illumina RNAseq. **(D)-(F)** For yeast (12 hours) treatment, the ONT library was compared to each of the libraries sequenced by Illumina RNAseq. The TE expression levels were given as reads per million mapped reads (RPM) for both sequencing methods. Spearman's correlation coefficient ρ was given per comparison, with each point representing expression levels of a TE family (x-axes=ONT and y-axes=Illumina).

### 6.4.2 Collection of expressed TEs

While Illumina sequencing outperforms the ONT platform in sequencing error rate, long-read technology surpasses the short-read sequencing method at the ability to resolve the complexities in TE-rich genomic regions. However, there's no standard method developed for the identification of active TEs using ONT data yet, nor an equation to define an ONT reads with a certain amount of Illumina reads. For instance, a 2kb ONT long read may equal 10+ Illumina reads. Therefore, it is reasonable to collect a conserved set of transcriptionally active TEs supported by two independent experiments and sequencing platforms. In the ONT library of mock treatment, 33,516 annotated TE loci overlapped with at least one ONT reads, of which 2,797 loci were identified as expression candidates in the Illumina data (Figure 6.4 A). These 2,797 TE loci identified by both sequencing platforms generally showed higher expression level (blue dots in Figure 6.4 B) in terms of ONT read counts and breadth of coverage than the 30,719 TE loci that only detected by ONT sequencing (grey

dots in Figure 6.4 B). Of the 2,797 TE loci identified by both ONT and Illumina sequencing, the proportion of the TE feature covered by ONT read in the overlapping pool clearly tilts toward 1 ('1' denotes 100% breadth of coverage), whereas the distribution of this proportion in TE loci uniquely identified by ONT platform has an additional bump close to 0 ('0' denotes 0% breadth of coverage; Figure 6.4 B). The detected TE loci in the yeast ONT library, in which 1,393 TE loci were supported by both sequencing platforms, show similar characteristics (Figure 6.5). These findings mean that the union of Illumina and ONT dataset refines the collection of transcriptionally active TEs in terms of the number of overlapping reads and breadth of coverage. Therefore these TE loci were considered as expressed TEs.

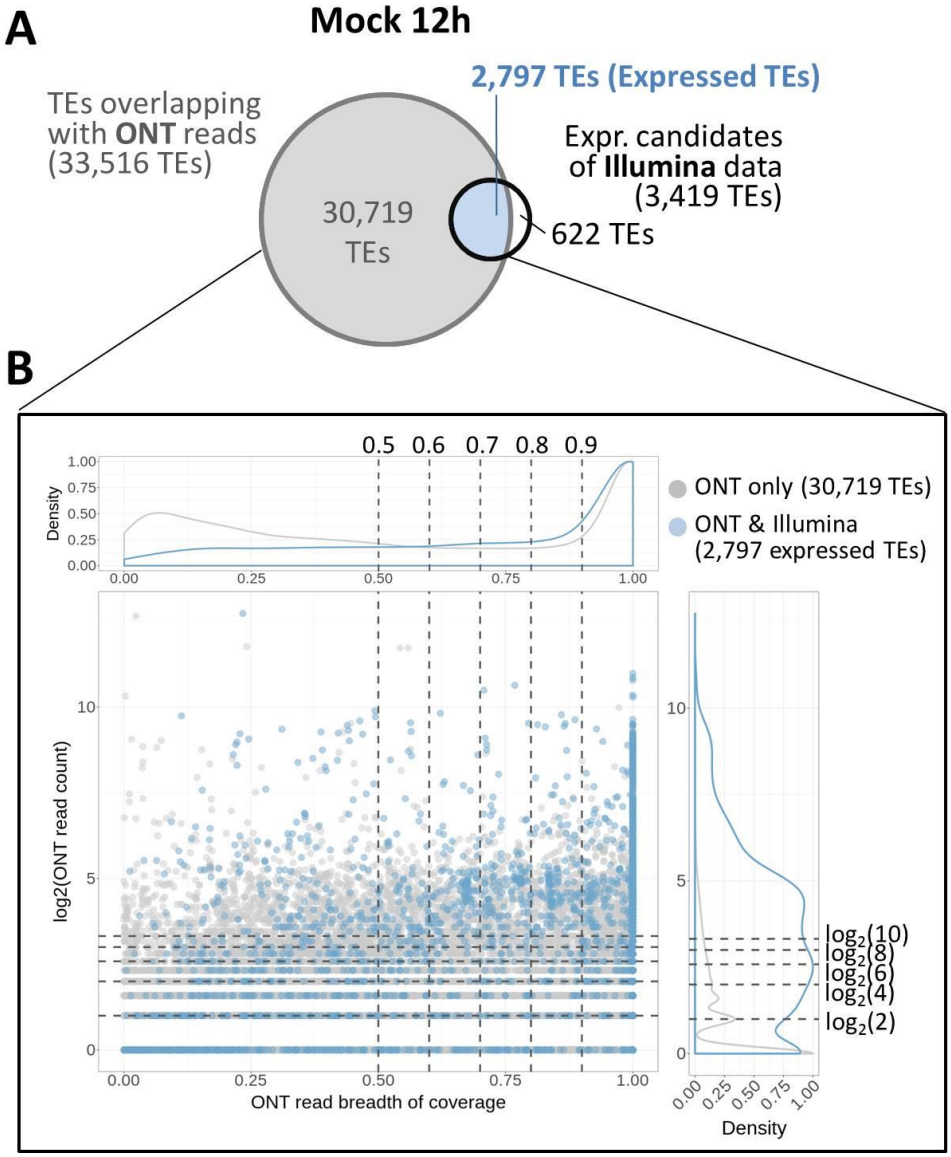



**Figure 6.4 Expressed TEs of mock treatment (12h)**

**(A) Collection of expressed TEs.** Expressed TE loci were identified by overlapping expression candidates (expr. candidates) of Illumina data and TE loci overlapping with at least one ONT read. TE loci of the inner joined region (blue) were considered expressed TEs. **(B) Expression range of TEs having at least one ONT read.** Grey dots denote TEs only appeared in ONT data, and blue dots denote expressed TEs supported by both ONT and Illumina data. The fraction of the TE feature

covered by ONT reads (x-axis) was plotted against the logarithmic-transformed read count (y-axis). The distributions of these two variables were plotted as density plots.

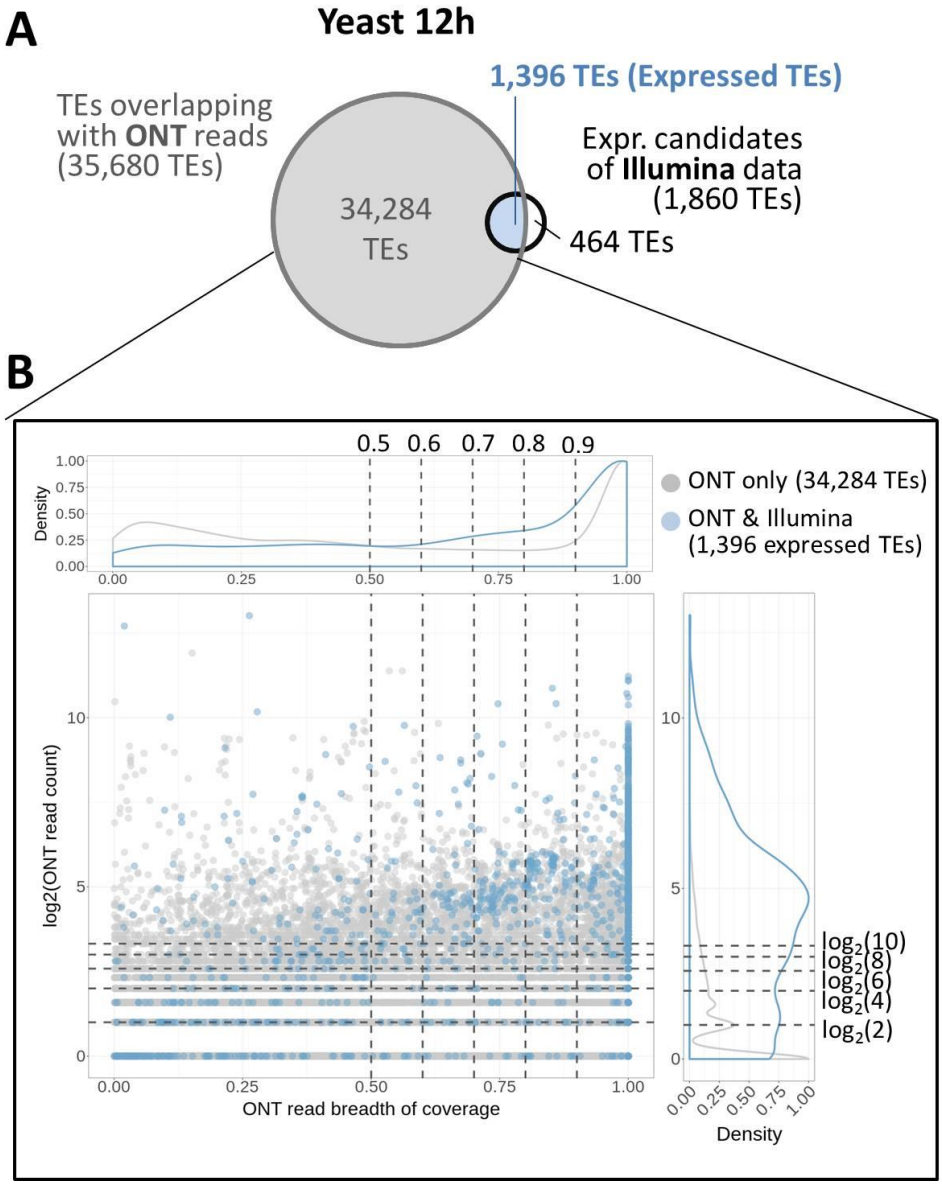



**Figure 6.5 Expressed TEs of yeast treatment (12h)**

**(A) Collection of expressed TEs.** Expressed TE loci were identified by overlapping expression candidates (expr. candidates) of Illumina data and TE loci overlapping with at least one ONT read. TE loci of the inner joined region were considered expressed TEs. **(B) Expression range of TEs having at least one ONT read.** Grey dots denote TEs only appeared in ONT data, and blue dots denote expressed TEs supported by both ONT and Illumina data. The fraction of the TE feature covered by ONT reads (x-axis) was plotted against the logarithmic-transformed read count (y-axis). The distributions of these two variables were plotted as density plots.

### 6.4.3 Validation of location bias and negative correlation of TE insertions and gene expression level

In chapter 3, expression candidates predominantly identified as being derived from TEs located in the intron of expressed genes (Figure 3.12). With additional support from ONT data, the expressed TEs captured in the previous section also show a similar proportion of transcriptionally active TEs co-localized with genes, particularly in the intron of expressed genes (Figure 6.6).

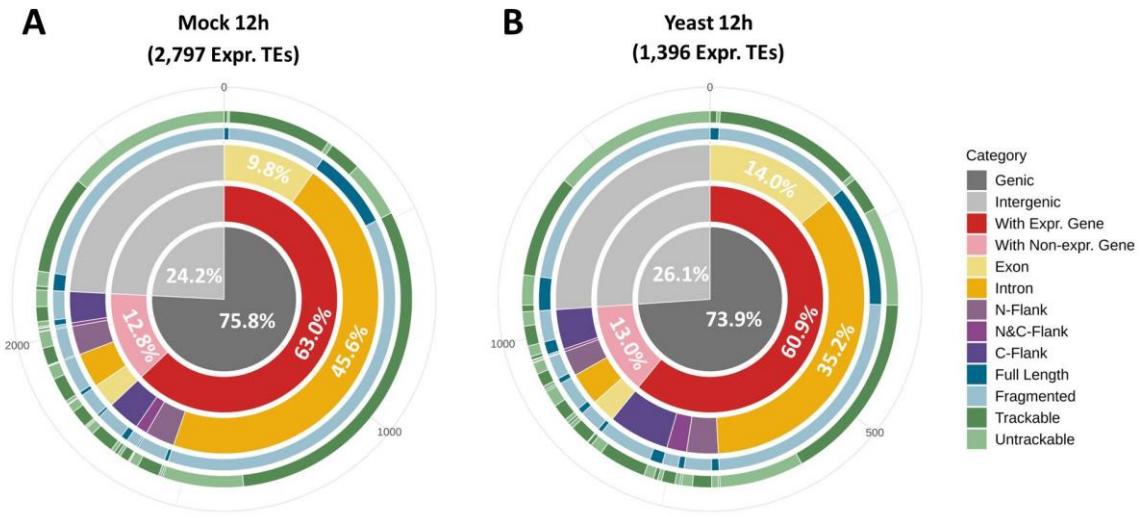



**Figure 6.6 Hierarchical classifications of expression candidates by location, integrity, and distinctness.**

Expressed TEs of **(A)** mock treatment (12h) and **(B)** yeast treatment (12h) were categorized in the order of region (centre), the transcriptional activity of co-localized genes (2nd layer), location (3rd layer), integrity (4th layer), and the presence/absence of unique-mapping reads (outer-most layer).

When considering gene transcription, the ONT and Illumina datasets were highly consistent in genes with TPM > 1 in the former and FPKM > 1 in the latter (Figure 6.7). In mock, 86.8% (13,326 genes) of genes showed TPM > 1 in ONT data are also included as expressed genes with FPKM > 1 in Illumina dataset (Figure 6.7 A). In yeast, this proportion dropped slightly to 70.3% (10,740 genes; Figure 6.7 B). Note that the numbers of genes with ONT TPM > 1 are similar in mock and yeast libraries, so the lower proportion of consistency in the yeast library might be due to fewer expressed genes (FPKM > 1) in the Illumina yeast dataset. In Table 3.2, the Illumina sequencing depth of yeast libraries was lower than other libraries, possibly due to the poorer RNA quality associated with yeast treatments. This may have led to the identification of an apparently lower number of expressed genes collected from yeast Illumina libraries than that from mock. Therefore the number of overlapping genes identified by both platforms in the yeast treatment is less than that in mock-treatment.

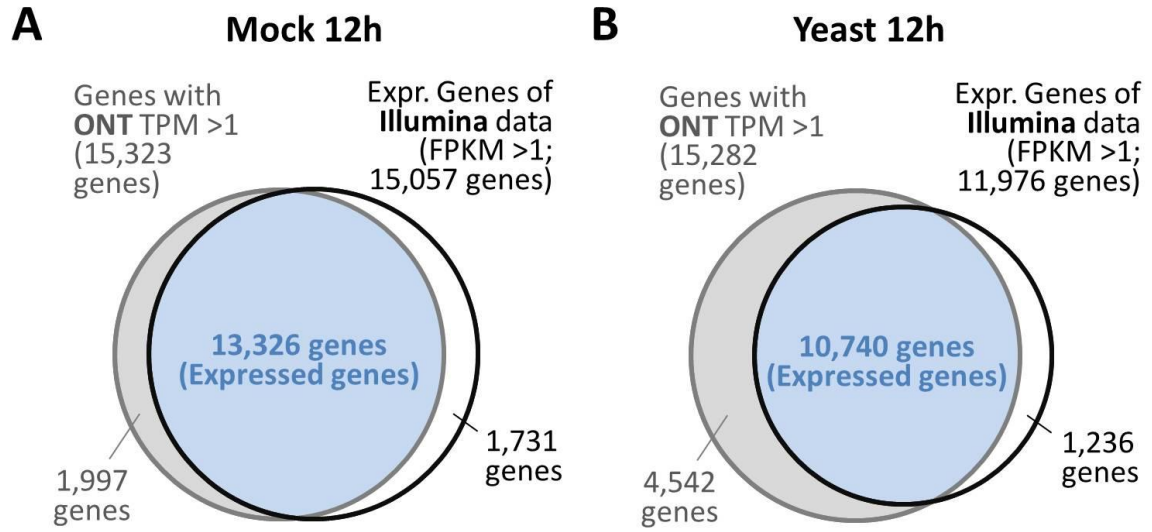



**Figure 6.7 Collection of expressed genes**

Genes with transcriptional activity in **(A)** mock treatment (12h) and **(B)** yeast treatment (12h) were identified by overlapping expressed genes (expr. genes) of Illumina data and genes with ONT transcripts per million (TPM) > 1. Genes of the inner joined region (blue) were supported by both sequencing platforms and therefore considered as expressed genes.

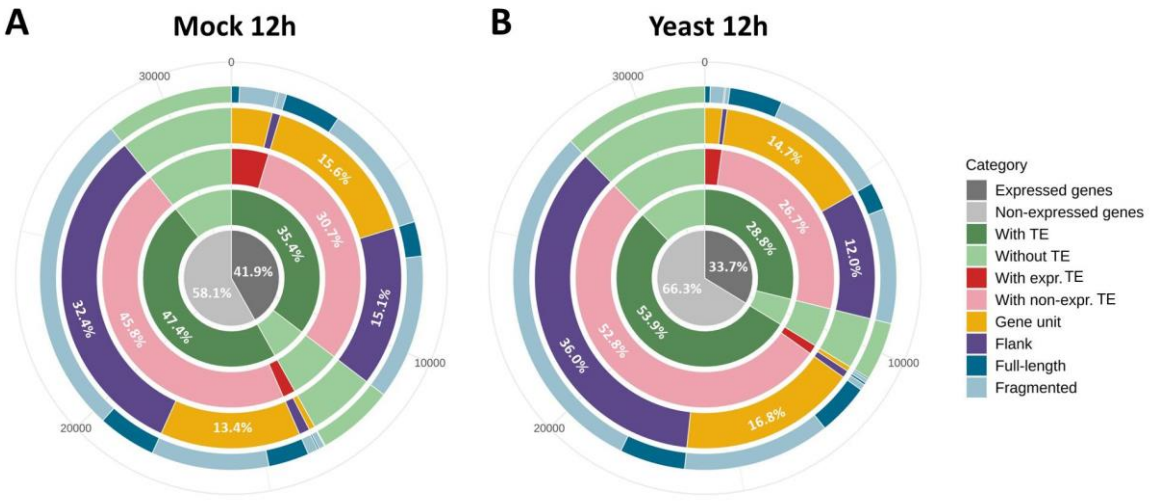



**Figure 6.8 Hierarchical categorization of all annotated genes.**

From the inner-most layer of the graph, all annotated genes were categorized by gene activity, presence/absence of TE insertions, presence/absence of expressed TEs, TE location, and TE integrity.

Consistent with the findings at T=0 (Figure 4.1), high proportions of expressed genes were found co-localized with TEs in the ONT dataset of mock- and yeast-treatment (Figure 6.8). Among genes with TPM > 1, genes only having inactive TEs in the gene body showed significantly lower expression level than genes without TEs (Figure 6.9 A, B, Figure 6.10 A, B), whereas genes having inactive TEs in flanking regions generally didn't show this negative correlation (Figure 6.9 C, D, Figure 6.10 C, D). On

the other hand, irrespective of the location of TE insertion, genes co-localized with expressed full-length TEs revealed significantly lower TPM than genes without TEs (Figure 6.9 E, G, Figure 6.10 E, G), while the expression level of genes with fragmented expressed TEs was not significantly deviated from that of genes without TEs (Figure 6.9 F, H, Figure 6.10 F, H). Overall, the observations obtained from the ONT dataset are concordant with that in Illumina data (section 4.4.1).

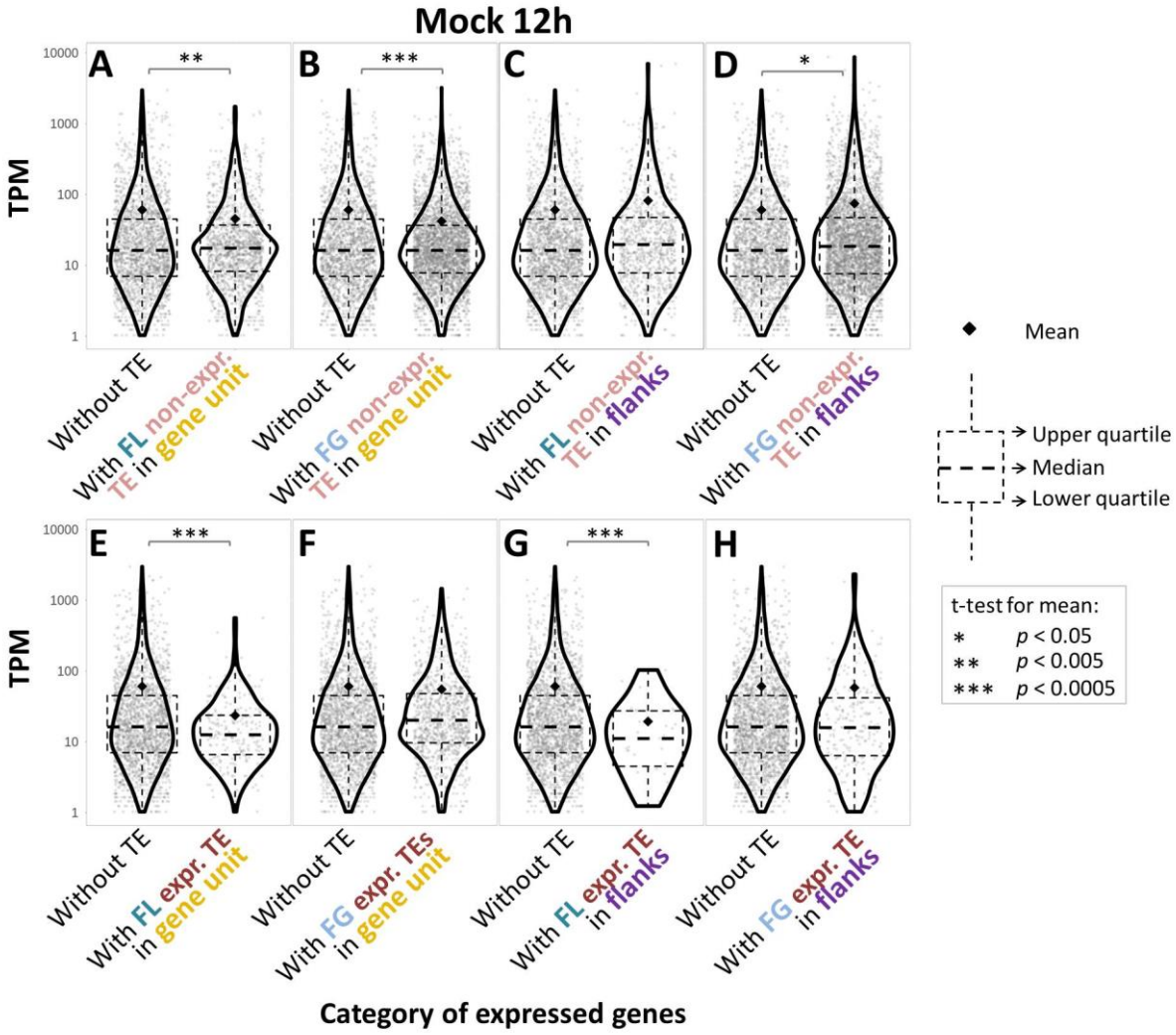



**Figure 6.9 Comparison of the expression level between genes without TE and those with TEs**

The TPM value of expressed genes without TE was compared pair-wise with expressed genes having **(A)** full-length (FL) expressed TEs in gene unit, **(B)** fragmented (FG) expressed TEs in gene unit, **(C)** FL expressed TEs in flanks, **(D)** FG expressed TEs in flanks, **(E)** FL non-expressed TEs in gene unit, **(F)** FG non-expressed TEs in gene unit, **(G)** FL non-expressed TEs in flanks, and **(H)** FG non-expressed TEs in flanks.

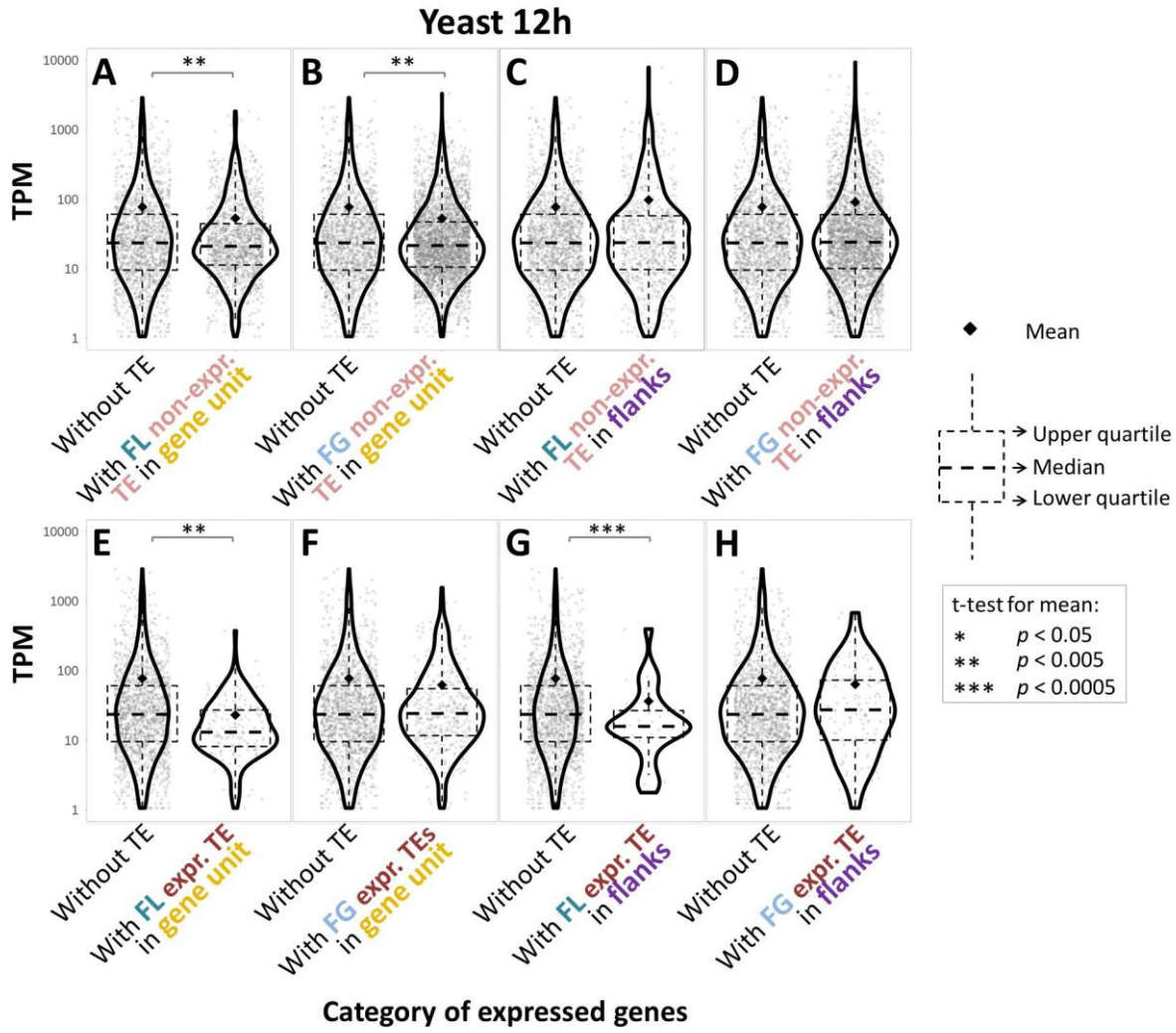



**Figure 6.10 Comparison of the expression level between genes without TE and those with TEs**

The TPM value of expressed genes without TE was compared pair-wise with expressed genes having **(A)** full-length (FL) expressed TEs in gene unit, **(B)** fragmented (FG) expressed TEs in gene unit, **(C)** FL expressed TEs in flanks, **(D)** FG expressed TEs in flanks, **(E)** FL non-expressed TEs in gene unit, **(F)** FG non-expressed TEs in gene unit, **(G)** FL non-expressed TEs in flanks, and **(H)** FG non-expressed TEs in flanks.

### 6.4.4 Alternative splicing associated with TEs

Although analysis of gene's TPM level reveals that transcriptionally active full-length TEs in gene units were less likely to associate with highly expressed genes than fragmented expressed TEs within genes, this analysis cannot tell whether active TEs, intact or fragmented, correlated with alternative splicing of genes. To investigate this, the FLAIR pipeline (Tang et al., 2020) was used to categorize alternative splicing events into four categories: alternative 3' splicing (Alt3), alternative 5' splicing (Alt5), intron retention (IR) and exon skipping (ES). Gene-related alternative splicing features overlapping with TEs were further collected, and the productivity (as per the definition in FLAIR pipeline, this denotes the ability of a transcript to produce protein) of gene transcripts having these

alternative splicing features was estimated. First of all, among the total 21,081 alternative splicing features identified by FLAIR across the ONT libraries of mock and yeast treatments, 19,526 (92.6 %) of them related to annotated genes. Over 90% of these gene-related alternative splicing features were IR (8,806 alternative splicing features) and ES (9,378 alternative splicing features). Note that an isoform may contain multiple numbers and various types of alternative splicing features. Nonetheless, an alternative splicing feature could appear in multiple isoforms, as indicated in Figure 6.13 A. Because a gene could have multiple isoforms, it is reasonable that the number of involved genes in each alternative splicing category is lower than the number of gene-related alternative splicing features. Notably, there are more genes than the number of associated ES features, suggesting that, for some ES events, each may involve more than one gene. Of the 19,526 gene-related alternative splicing features, only 524 (2.7%) of them overlapping with TEs (Figure 6.13 A). As expected, almost all TEs overlapping with Alt3, Alt5 and IR features are located within introns, while 22 of 40 ES-associated TEs overlapped with annotated exons.

**A**

| Annotation: | All AS | | Gene-related AS | | | Gene&TE-related AS | | | |
|---|---|---|---|---|---|---|---|---|---|
| | # of AS | # of isoforms | # of AS | # of isoforms | # of genes | # of AS | # of isoforms | # of genes | # of TEs |
| Alt3 | 900 | 2,647 | 815 | 2,454 | 697 | 125 | 229 | 108 | 267 |
| Alt5 | 576 | 1,815 | 527 | 1,683 | 486 | 72 | 117 | 69 | 168 |
| IR | 10,227 | 19,089 | 8,806 | 17,116 | 4,621 | 286 | 324 | 273 | 334 |
| ES | 9,378 | 15,796 | 9,378 | 15,796 | 13,943 | 41 | 64 | 35 | 40 |

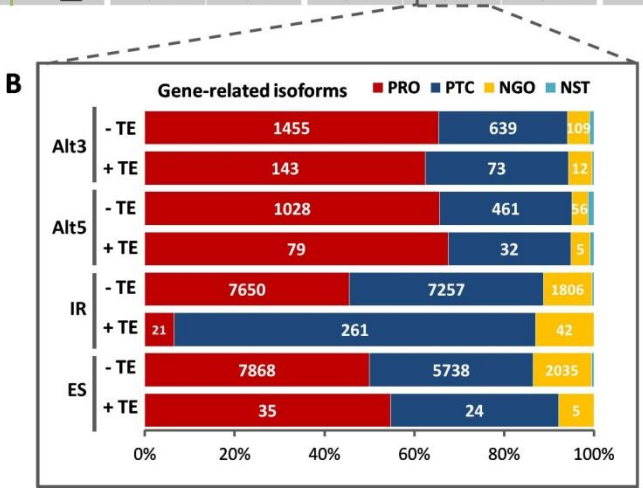



**Figure 6.11 Categorization of alternative splicing event**

**(A)** As illustrated in the first column, there are four types of alternative splicing (AS): alternative 3' splicing (Alt3), alternative 5' splicing (Alt5), intron retention (IR), and exon skipping (ES). The numbers of AS features, isoforms, associated genes and TEs were as indicated. **(B)** The productivity of gene-related isoforms can be categorized as productive (PRO), having premature termination codon (PTC), having no start codon (NGO), or having a start codon but no stop codon (NST). Isoforms containing AS features overlapping with TEs were labelled as "+TE"; otherwise labelled as "-TE".

To understand whether the presence of these TEs associated with the productivity of the gene transcripts, the productivity of isoforms containing these gene-and-TE-related alternative splicing was estimated using FLAIR and grouped into four types, productive (PRO), having premature termination codon (PTC, i.e. unproductive), no start codon (NGO), and having start codon but no stop codon (NST). This analysis shows that 50 % to 68 % of the isoforms having Alt3, Alt5, or ES remained productive, no matter whether the alternative splicing features overlapped with TEs (Figure 6.13 B). However, 80.6 % of isoforms having TE-related IR were PTC, while the PTC proportion in isoforms having IR events non-overlapping with TEs was less than 45%. Looking into the estimated translation stop site of these isoforms containing TE-related IR feature, 196 of the 261 PTC isoforms exhibit premature stop codon exactly within the TE-overlapping IR feature. From the perspective of the isoform orientation, nine of the translational premature termination sites appear within TEs, two are after TEs, and the rest 186 isoforms show premature termination sites before the presence of TEs. The distance between TEs and the premature termination sites presented prior to TEs ranged from 2 bp to over 4 kb, with the first quartile, median and third quartile at 147 bp, 311 bp and 693 bp, respectively.

Interestingly, different TE superfamilies were preferentially observed among the four types of alternative splicing features. Retrotransposon VLINE was over-represented in Alt3 and Alt5 alternative splicing events (Figure 6.14 A, B), and Harbinger, a DNA transposon, was predominantly seen in IR features (Figure 6.14 C). For ES features, MULE DNA transposon was the most predominant superfamily among all TE superfamily (Figure 6.14 D). In addition, most of these TEs were fragmented. There’re only 10, 7, 34 and 1 full-length TE loci associated with Alt3, Alt5, IR, and ES features, respectively.

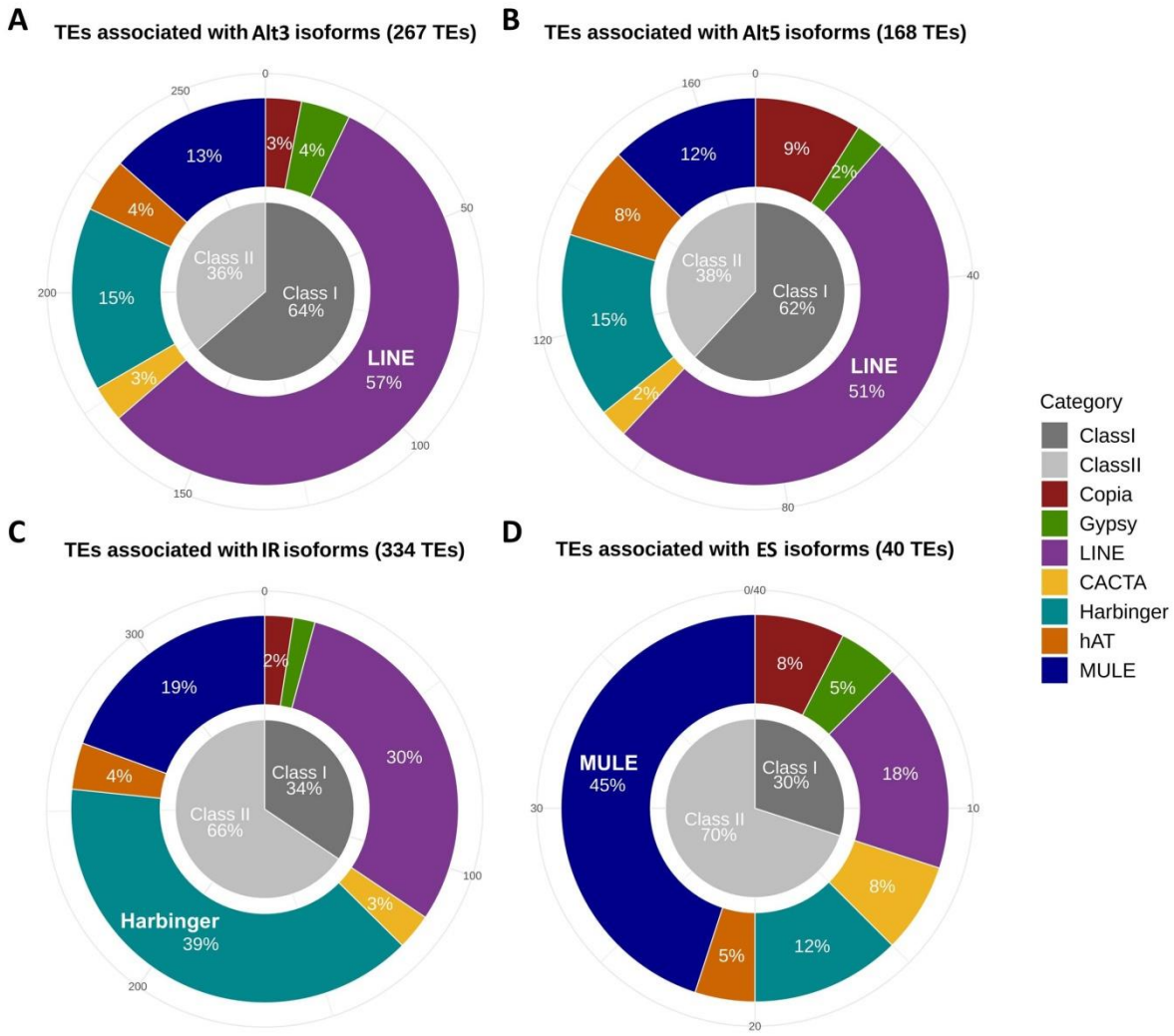



**Figure 6.12 Categorization of TEs associated with alternative splicing**

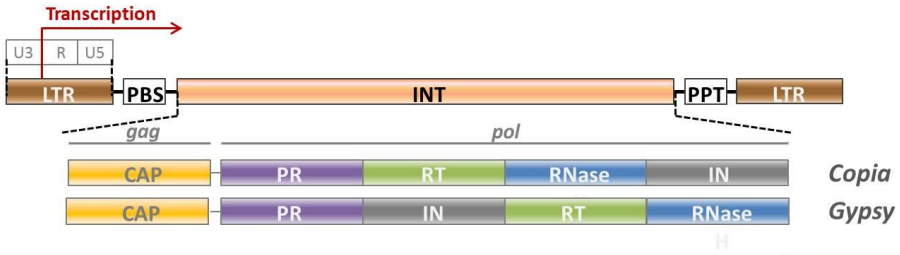
TE loci overlapping with gene-related **(A)** Alt3, **(B)** Alt5, **(C)** IR, and **(D)** ES features were grouped by class (central pie graph), then by superfamily (outer doughnut graph). The percentage of each slice is as indicated, with the most over-represented superfamily labelled.

### 6.4.5 Identification of autonomous TEs having full transcription

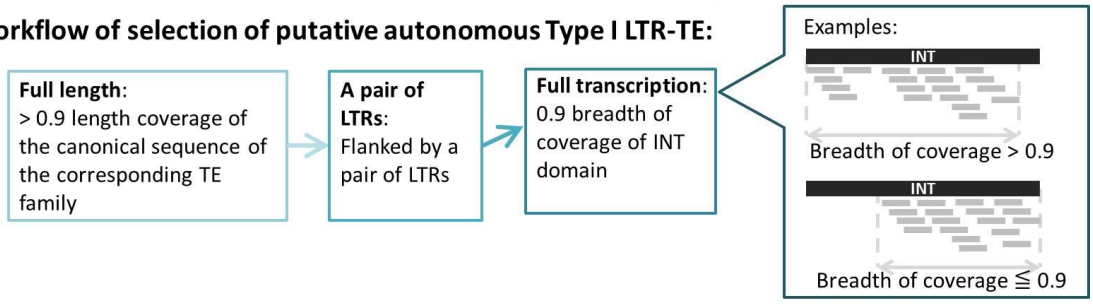
Due to the ability of ONT sequencing technology to sequence full-length transcripts, ONT cDNA libraries have the potential to reveal, if any, competent transcription of autonomous TE and decipher the origins of these transcripts. Using the workflow established in chapter 4, intact LTR-TEs with >90% INT coverage (Figure 6.15 A, B), autonomous LINEs with >0.9 breadth of coverage across whole elements (Figure 6.16 A, B), as well as intact TIR-transposon with >90% ORF covered by ONT reads (Figure 6.17 A, B) were collected. This process captured 20 and 19 LTR-TE loci in mock and yeast libraries, respectively (Figure 6.15 C). These include Copia-3, Copia-23, and Gypsy-V1. For LINE retrotransposon, only a single VLINE7 locus and two VLINE 8 loci were selected from the mock library

(Figure 6.16 C). For TIR-TE, three hAT-7 loci in the mock library revealed >0.9 breadth of coverage across ORF (Figure 6.17 C).

**A** **Structure of Type I LTR retrotransposon:**




**B** **Workflow of selection of putative autonomous Type I LTR-TE:**



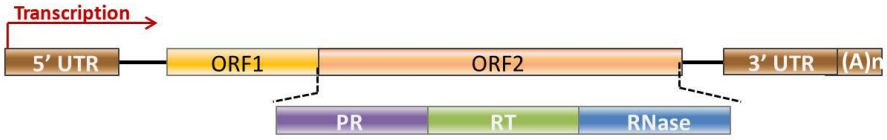

**C** **Investigation of putative autonomous Type I LTR-TE:**

| | # Annotated TEs | # Intact TEs | # Full transcription | |
|---|---|---|---|---|
| | | | Mock 12h | Yeast 12h |
| Copia | 44,598 | 1,182 | Copia-23 : 19 | Copia-3: 8<br>Copia-23: 11 |
| Gypsy | 64,827 | 498 | Gypsy-V1: 1 | 0 |
| Sum | 109,425 | 1,680 | 20 | 19 |

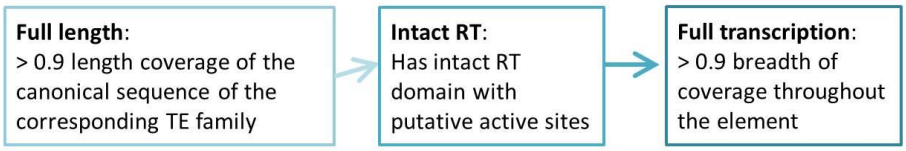
**Figure 6.13 Identification of autonomous LTR-TE with potential full-transcription.**

**(A)** Autonomous LTR-TEs are characterized by the poly-protein-coding internal domain (INT) flanked by a pair of long terminal repeats (LTR). The transcription starts within the 5' LTR, going through the INT domain, and generally terminated at 3' LTR. See Figure 4.1 for the acronyms. The diagram is not drawn to scale. **(B)** Workflow for collecting autonomous LTR-TE loci that were potentially fully transcribed. The short grey segments denote sequencing reads. **(C)** The numbers of annotated, intact, and potentially fully transcribed LTR-TE loci.

**A** **Structure of Type I LINE retrotransposon:**




**B** **Workflow of selection of putative autonomous Type I LINE:**

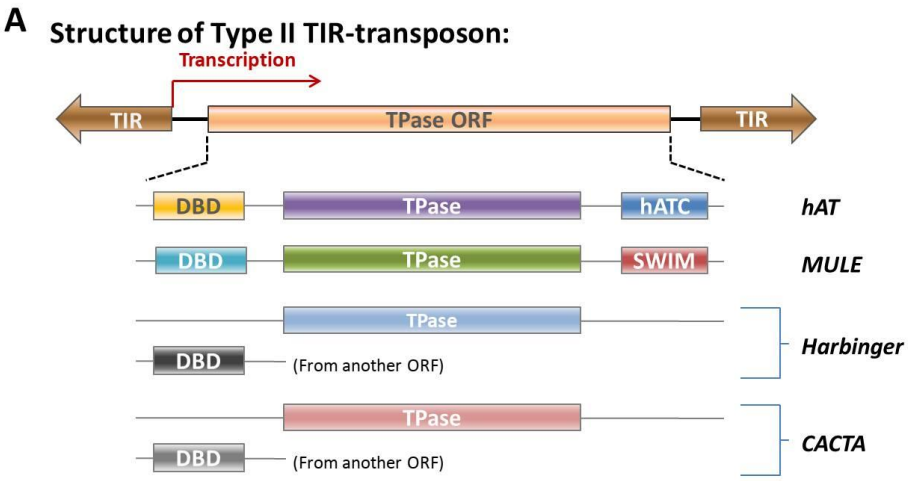


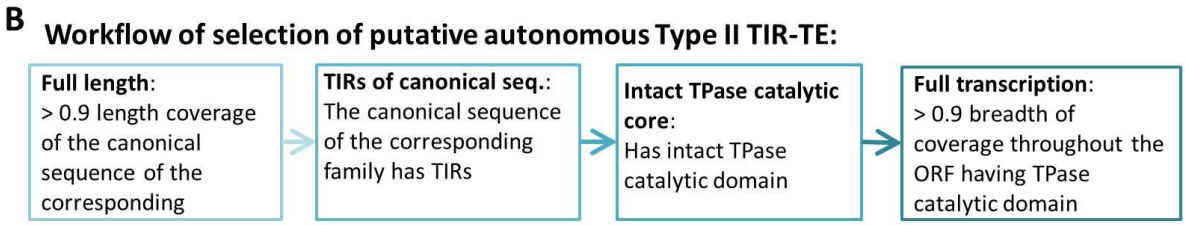

**C** **Investigation of putative autonomous Type I LINE:**

| | # Annotated TEs | # Intact TEs | # Full transcription | |
|---|---|---|---|---|
| | | | Mock 12h | Yeast 12h |
| LINE | 23,447 | 122 | VLINE7: 1<br>VLINE8: 2 | 0 |
| Sum | 23,447 | 122 | 3 | 0 |

**Figure 6.14** **Identification of autonomous LINE with potential full-transcription.**

**(A)** An autonomous LINE is expected to retain the open reading frames (ORF) encoding proteins necessary for mobilization. Typically, the transcription starts from the 5' untranslated region (5' UTR), going through ORFs and 3' UTR, and completed with polyadenylation. See Figure 4.2 for the acronyms. The diagram is not drawn to scale. **(B)** Workflow for collecting autonomous LINE loci that were potentially fully transcribed. **(C)** The numbers of annotated, intact, and potentially fully transcribed LINE loci.




C **Investigation of putative autonomous Type II TIR-TE :**

| | # Annotated TEs | # Intact TEs | # Full transcription | |
|---|---|---|---|---|
| | | | Mock 12h | Yeast 12h |
| hAT | 15,374 | 170 | hAT-7: 3 | 0 |
| MULE | 27,336 | 25 | 0 | 0 |
| Harbinger | 32,053 | 3 | 0 | 0 |
| CACTA | 12,632 | 10 | 0 | 0 |
| Sum | 87,395 | 208 | 3 | 0 |

**Figure 6.15 Identification of autonomous TIR-TEs with potential full-transcription.**

**(A)** An autonomous TIR-TE is supposedly equipped with a transposase (TPase)-encoding ORF flanked by terminal inverted repeats (TIR). The typical transcription start site is as indicated. See Figure 4.3 for the acronyms. The diagram is not drawn to scale. **(B)** Workflow for collecting autonomous TIR-TE loci that were potentially fully transcribed. **(C)** The number of annotated, intact, and potentially fully transcribed TIR-TE loci.

To further check whether the nearly full breadth of coverage resulted from contiguous full-length ONT reads across TEs, rather than a co-contribution of multiple reads, the read length of each read and the bases mapped to the potentially autonomous TEs were investigated. Proving that a full-length autonomous TE transcript was present, the resulting ONT read should meet two criteria. Firstly, depending on the type of mapped TE, it should be at least as long as the INT domain, the ORF, or the full feature of the TE locus. Secondly, this read should have its TE-mapped bases almost as much as its read length. For example, the ONT reads mapping to the 19 seemly fully expressed

autonomous Copia-23 loci in the mock library were mostly shorter than 3,000 bp (x-axis, Figure 6.18 A), whereas the size of the canonical Copia-23 INT domain is 4,084 bp. The only read longer than 3 kb identified exhibited a very poor mapping to the element (y-axis, Figure 6.18 A). In addition, the majority of these ONT reads were multi-mapping (red dots, Figure 6.18 A). This analysis showed that most of these reads were skewed from the diagonal line, indicating the inconsistency between the lengths of ONT reads and the mapped bases of these reads to TEs. To figure out the factors underlying this inconsistency, the alignment start and end sites of these ONT reads in relation to the mapped Copia-23 loci were surveyed. As illustrated in the cartoons in Figure 6.18 B and C, the head and tail of the alignment were grouped into three categories, internal, external, and clipped. The investigation reveals that most of these ONT reads represented transcription started within the Copia-23 loci (Figure 6.14 B). However, the tail of the reads, especially those that deviated from the diagonal line, were mostly clipped due to the sequence discrepancy between ONT reads and TEs (Figure 6.18 C). Only a few of them extended through the annotated boundary. Overall, there is no evidence of autonomous transcription from annotated Copia-23 loci.

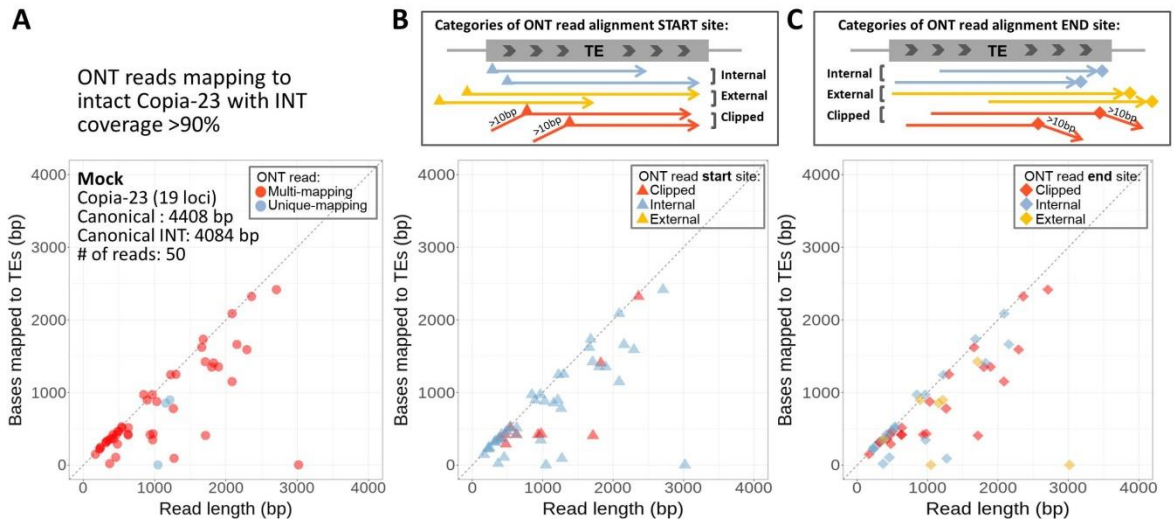



**Figure 6.16 Characteristics of ONT reads mapping to autonomous Copia-23 loci identified by the workflow shown in Figure 6.15 in mock treatment.**

**(A)** To identify a single long read representing full transcription of the autonomous TE loci, the read length was plotted against the number of bases overlapping with the autonomous locus. The diagonal dashed line denotes where length equals mapped bases. Red spots are multi-mapping reads; blue spots are unique-mapping reads. **(B-C)** To understand why many ONT reads only partially aligned with TE loci, all reads plotted in (A) were presented at the same coordinates in (B) and (C), yet coloured by the types of alignment start sites (B) and end sites (C) relative to TEs. Colours in (B) and (C): red denotes clipped read; blue denotes alignment started or ended internally; yellow denotes alignment started or ended externally.

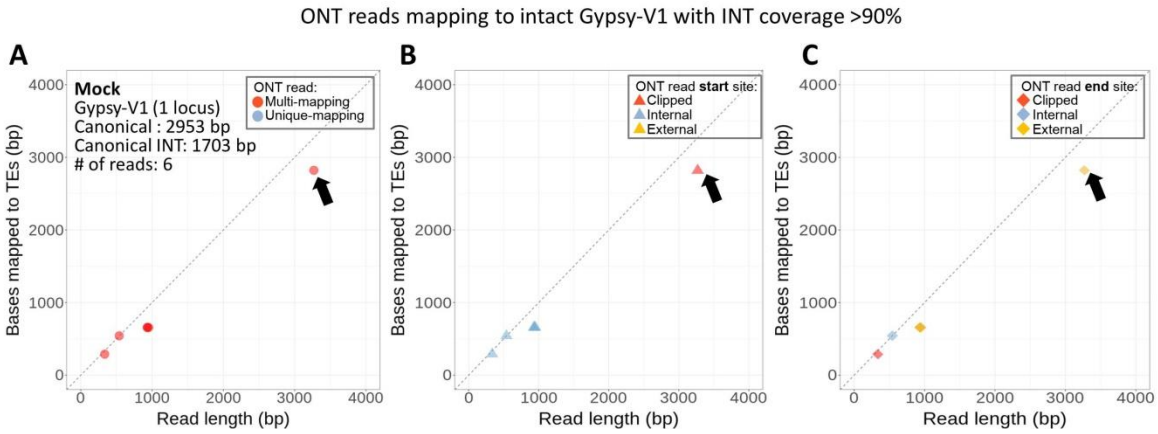



**Figure 6.17 Characteristics of ONT reads mapping to autonomous Gypsy-V1 locus identified by the workflow shown in Figure 6.15, in mock treatment.**

**(A)** The read length was plotted against the number of bases overlapping with the autonomous locus. The diagonal dashed line denotes where length equals mapped bases. Red spots are multi-mapping reads; blue spots are unique-mapping reads. The black arrow indicates a long read that covered the majority of this autonomous TE locus. **(B-C)** All reads plotted in (A) were presented at the same coordinates in (B) and (C), yet coloured by the types of alignment start sites (B) and end sites (C) relative to TEs. Colours in (B) and (C): red denotes clipped read; blue denotes alignment started or ended internally; yellow denotes alignment started or ended externally.

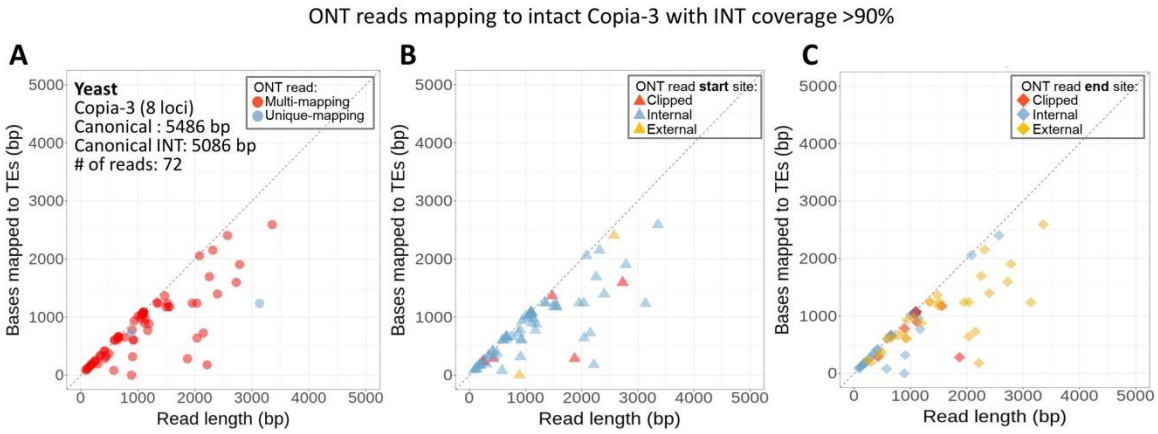



**Figure 6.18 Characteristics of ONT reads mapping to autonomous Copia-3 loci identified by the workflow shown in Figure 6.15, in yeast treatment.**

**(A)** The read length was plotted against the number of bases overlapping with the autonomous loci. The diagonal dashed line denotes where length equals mapped bases. Red spots are multi-mapping reads; blue spots are unique-mapping reads. **(B-C)** All reads plotted in (A) were presented at the same coordinates in (B) and (C), yet coloured by the types of alignment start sites (B) and end sites (C) relative to TEs. Colours in (B) and (C): red denotes clipped read; blue denotes alignment started or ended internally; yellow denotes alignment started or ended externally.

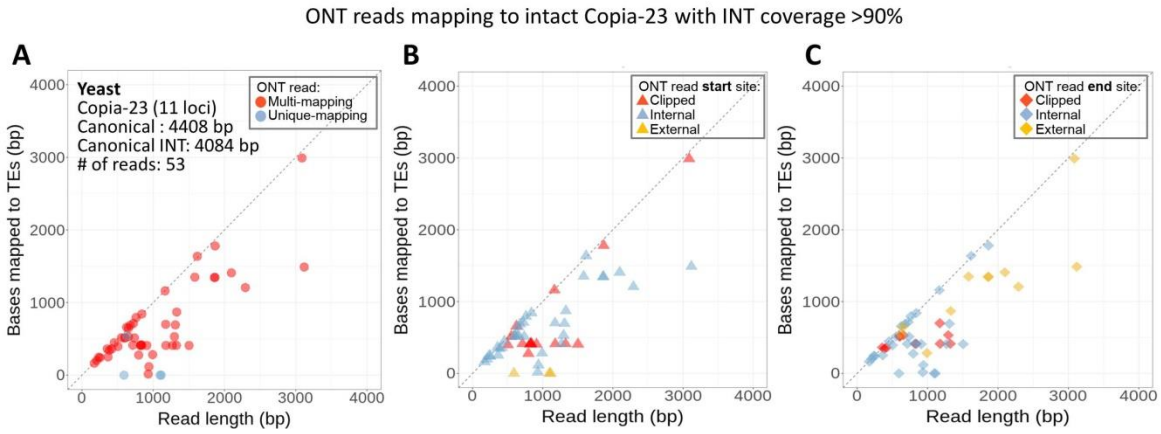



**Figure 6.19 Characteristics of ONT reads mapping to autonomous Copia-23 loci identified by the workflow shown in Figure 6.15 in yeast treatment.**

**(A)** The read length was plotted against the number of bases overlapping with the autonomous loci. The diagonal dashed line denotes where length equals mapped bases. Red spots are multi-mapping reads; blue spots are unique-mapping reads. **(B-C)** All reads plotted in (A) were presented at the same coordinates in (B) and (C), yet coloured by the types of alignment start sites (B) and end sites (C) relative to TEs. Colours in (B) and (C): red denotes clipped read; blue denotes alignment started or ended internally; yellow denotes alignment started or ended externally.

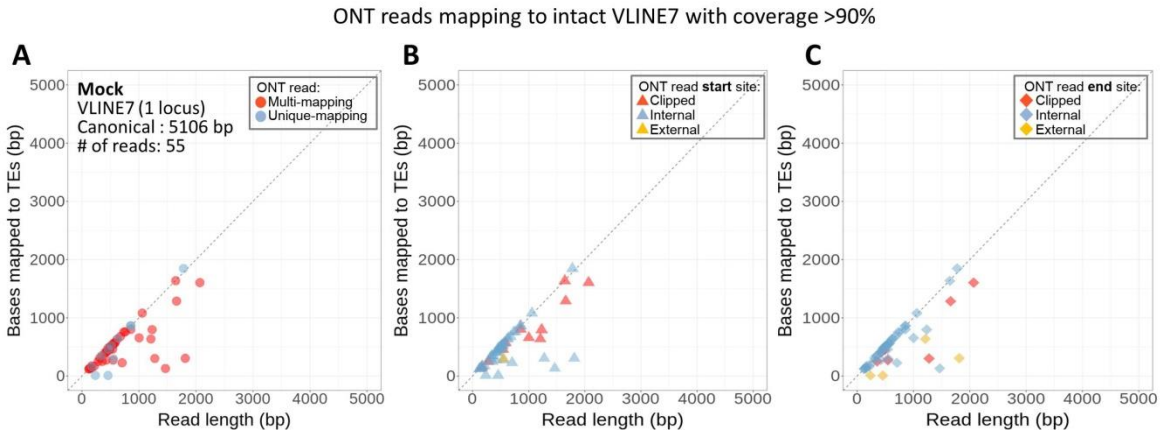



**Figure 6.20 Characteristics of ONT reads mapping to autonomous LINE7 locus identified by the workflow shown in Figure 6.16 in mock treatment.**

**(A)** The read length was plotted against the number of bases overlapping with the autonomous locus. The diagonal dashed line denotes where length equals mapped bases. Red spots are multi-mapping reads; blue spots are unique-mapping reads. **(B-C)** All reads plotted in (A) were presented at the same coordinates in (B) and (C), yet coloured by the types of alignment start sites (B) and end sites (C) relative to TEs. Colours in (B) and (C): red denotes clipped read; blue denotes alignment started or ended internally; yellow denotes alignment started or ended externally.

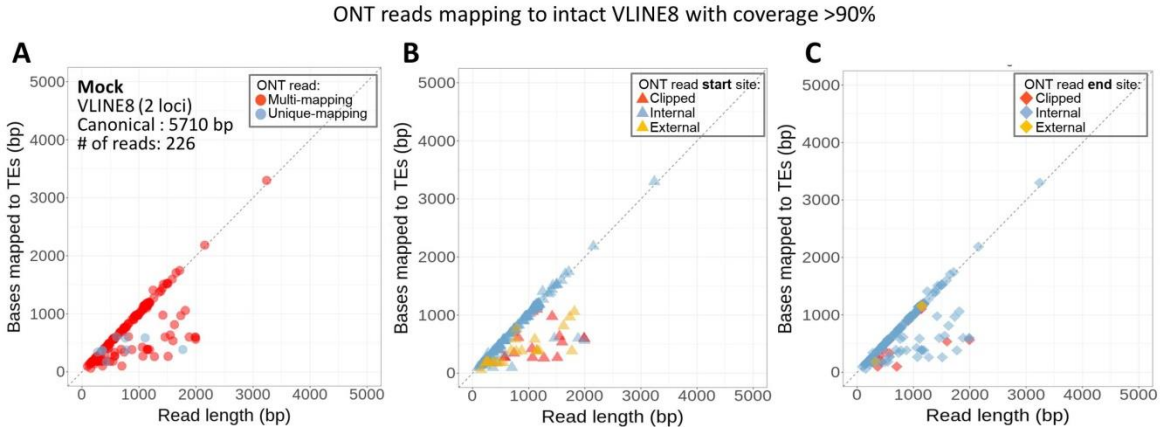



**Figure 6.21 Characteristics of ONT reads mapping to autonomous LINE8 loci identified by the workflow shown in Figure 6.16 in mock treatment.**

**(A)** The read length was plotted against the number of bases overlapping with the autonomous loci. The diagonal dashed line denotes where length equals mapped bases. Red spots are multi-mapping reads; blue spots are unique-mapping reads. **(B-C)** All reads plotted in (A) were presented at the same coordinates in (B) and (C), yet coloured by the types of alignment start sites (B) and end sites (C) relative to TEs. Colours in (B) and (C): red denotes clipped read; blue denotes alignment started or ended internally; yellow denotes alignment started or ended externally.

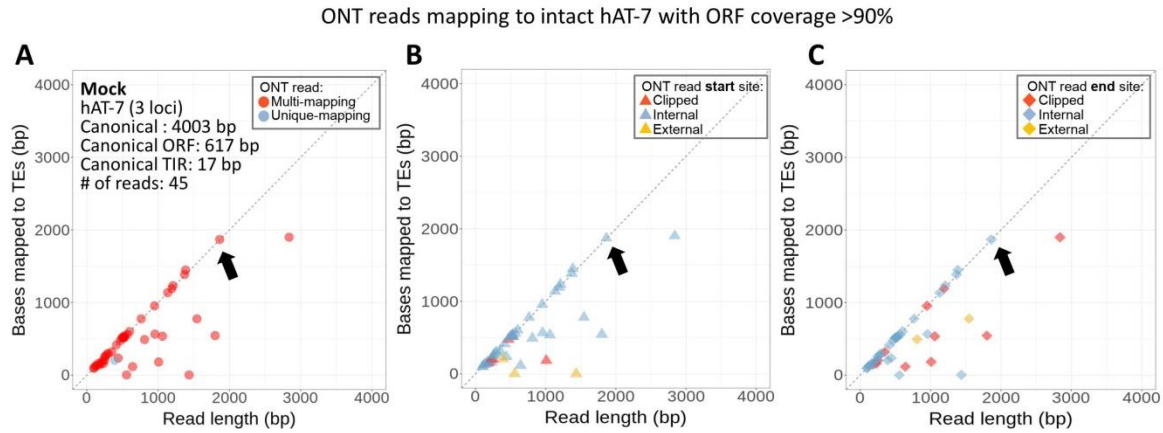



**Figure 6.22 Characteristics of ONT reads mapping to autonomous hAT-7 loci identified by the workflow shown in Figure 6.17 in mock treatment.**

**(A)** The read length was plotted against the number of bases overlapping with the autonomous loci. The diagonal dashed line denotes where length equals mapped bases. Red spots are multi-mapping reads; blue spots are unique-mapping reads. The black arrow indicates a long read that may have covered the majority of the ORF. **(B-C)** All reads plotted in (A) were presented at the same coordinates in (B) and (C), yet coloured by the types of alignment start sites (B) and end sites (C) relative to TEs. Colours in (B) and (C): red denotes clipped read; blue denotes alignment started or ended internally; yellow denotes alignment started or ended externally.

The situation described for the 19 autonomous Copia-23 loci was also observed in captured TE loci of Copia-3, Gypsy-V1, VLINE7, VLINE8, and hAT-7 (Figure 6.19 – Figure 6.24). Only a single ONT read mapping to the autonomous Gypsy-V1 loci, and eight reads of hAT-7 appeared to adequately cover the bases of the INT (Figure 6.19) or ORF (Figure 6.24) of the associated TE loci. The genome browser image of the only Gypsy-V1 locus demonstrates the full coverage of this locus by a single ONT read (Figure 6.25). The genome browser image for hAT-7 shows ONT reads covering the ORF of the hAT-7 in chromosome 14 (Figure 6.26). These suggest potential transcription of Gypsy-V1 and hAT-7 may allow limited mobilisation of these elements.

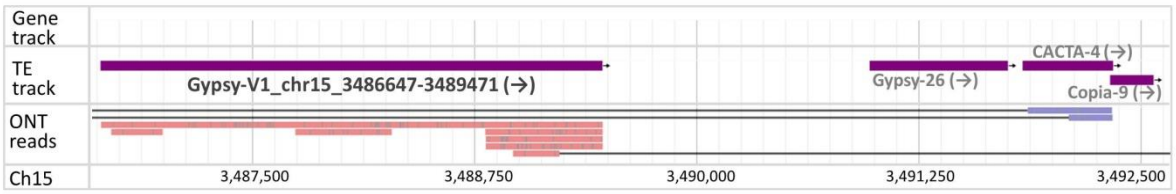



**Figure 6.23 Genome browser image of the autonomous Gypsy-V1 fully covered by ONT read.**

This TE locus, Gypsy-V1_chr15_3486647-3489471, is the only LTR-TE fully covered by a single ONT read. It locates in chromosome 15 and does not co-localize with any gene. The pink and blue strips denote forward and reverse reads, respectively. The orientations of annotated TEs were as indicated.

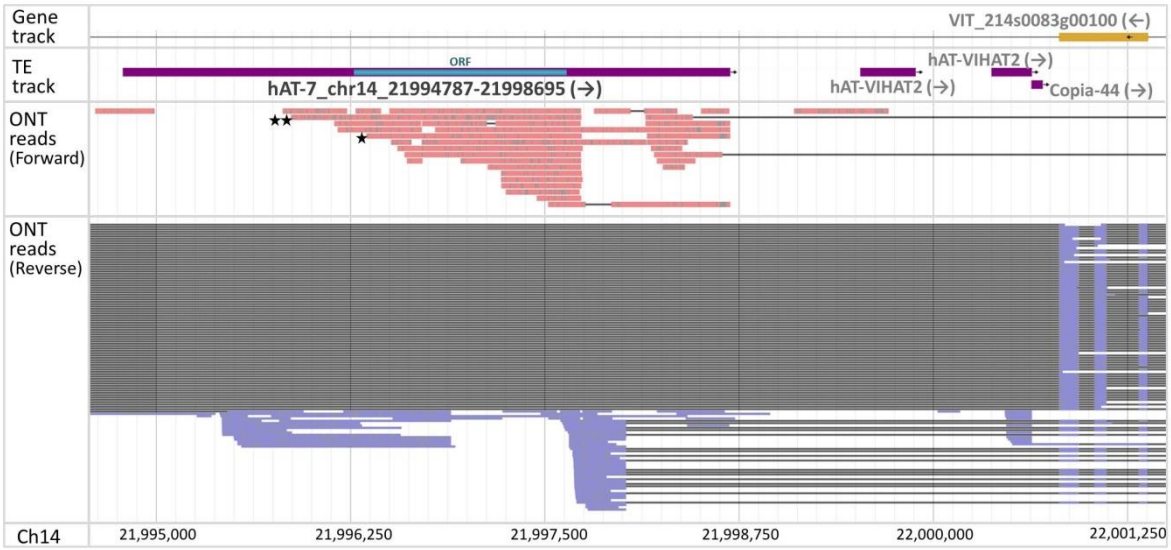



**Figure 6.24 Genome browser image of the representative autonomous hAT-7.**

This TE locus, hAT-7_chr14_21994787-21998695, is one of the three hAT-7 loci having ORF fully covered by individual ONT reads. The ORF (teal blue strip) identified using ORFfinder was overlaid manually and proportionally. It locates within an intron of VIT_214s0083g00100. The pink and blue strips denote forward and reverse reads, respectively. The two black stars label the ONT read fully covering the ORF, while the single black star marks the ONT read overlapping with >90% ORF in the sense orientation. The orientations of annotated TEs and genes were as indicated.

### 6.4.6 Stress-related CREs annotated in the LTRs of Copia-3, Copia-23, and Gypsy-V1

In the Illumina dataset, Copia-3 and Copia-23 were the two TE families showing the most prominent potential of producing autonomous transcripts in both mock and yeast treatment (see Chapter 3). However, using ONT cDNA sequencing, there is no evidence of full-length ONT reads mapping through any of the annotated autonomous Copia-3 and Copia-23 loci. By contrast, among the LTR-retrotransposon, there was one ONT read covering a potentially autonomous Gypsy-V1 locus in the mock-treatment dataset. The mock treatment involves vigorous physical manipulation of the callus, potentially representing a wound stress treatment of the embryogenic callus. Nevertheless, the lack of significant read depth in the ONT dataset suggests that the stressors applied to these calli may not be sufficient or the correct stimuli to activate TE transcription through their stress-responsive cis-regulatory elements (CRE).

The survey of the CRE within canonical LTRs reveals that there are two pathogen-related CREs in Copia-3's LTR (Figure 6.25 A). In the LTR of Copia-23, there are three elicitor-responsive, two pathogen-related, and two wound-responsive CREs (Figure 6.25 B). These may partially explain the transcriptional activation of Copia-3 and Copia-23 observed in both Illumina and ONT data. However, their scattered distribution may not sufficiently drive the initiation of the full transcription and thus mobilization potential of these elements. On the contrary, Gypsy-V1's LTR has 22 pathogen-related, six elicitor-responsive, and four wound-responsive elements (Figure 6.25 C). Furthermore, heat-related CREs are conspicuously present in the LTR of Gypsy-V1 and concentrated together as an island of heat-responsive elements.

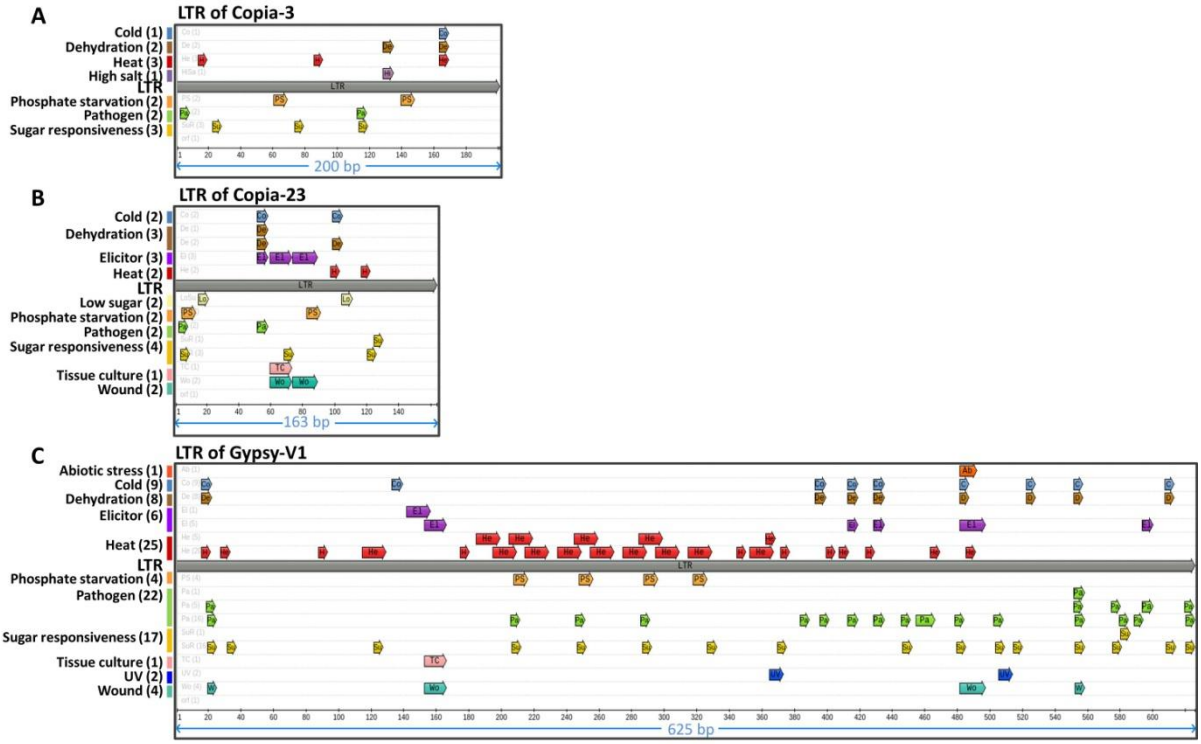



**Figure 6.25 Survey of stress-related CREs of the LTR of Copia-3, Copia-23 and Gypsy-V1**

The annotated CREs in the LTR of **(A)** Copia-3, **(B)** Copia-23, and **(C)** Gypsy-V1 were denoted by arrows in the genome browser. The stress stimuli were labelled at the left side with the number of corresponding CREs indicated in the brackets.

## 6.5 Discussion

### 6.5.1 Comparable average alignment depth and gene expression quantification between the ONT and Illumina libraries

The sequencing depth and length distribution of sequencing reads (indicated by N50) of long-read sequencing data are determinants for the detection of low-level transcription. These are especially important for TE-oriented study for that the transcriptional level of autonomous TEs might be low relative to the general expression level of genes. Therefore the comparability between the interrogated ONT cDNA sequencing dataset and other known datasets is of great importance.

In this chapter, the ONT cDNA sequencing generated raw sequencing output equal to 17.6X and 20.4X coverage of the *V. vinifera* reference genome (Figure 7.1). This coverage is higher than the 10X coverage in the Illumina libraries used in previous chapters. In addition, the ONT sequencing depth seems comparable with and proportional to the sequencing output of a recently reported *Arabidopsis* ONT cDNA sequencing data (Panda and Slokin, 2020), in which full-length transcripts of autonomous TEs were detected. The alignment depth of our ONT cDNA data is roughly equal to or higher than that of the Illumina libraries investigated in the previous chapters (Figure 7.1 B, E). The N50 values seemed to be at the lower end compared with the distribution of length of annotated genes (median = ~1.2kb) in the grapevine's reference genome. However, the N50 of cDNA sequencing might be various from one species to another. Data from published literature show N50 of 948 to 972 bp in *Arabidopsis* (Panda and Slokin, 2020), 1.2 to 1.7 kb in mouse (Sessegolo et al., 2019), and 771 bp in human (Workman et al., 2019).

Among ONT reads mapping to the reference genome, 118K to 139K of them overlapping with annotated TEs, contributing to 1.7% of the total mapped reads (Table 7.1). The proportion of TE-mapped reads obtained from our ONT data is similar to our Illumina dataset. Reads overlapping with TEs showed an elevation of N50 to 1.2 kb. However, we cannot rule out whether there's an under-representation of large TE transcripts. These long TE transcripts may have a fast turn-over rate or have been quickly processed by epigenetic components, e.g. DCL1, DCL2, DCL4 etc. (Cuerda-Gil and Slotkin, 2016). The kinetics of sequencing in an ONT flow cell ensures that smaller fragments sequence more efficiently than longer fragments due to increased molar ends of short fragments and shorter Nanopore occupation times. Size selection prior to ONT cDNA sequencing to induce a bias towards longer cDNAs may enrich long reads in the library for the purpose of capturing autonomous TE transcripts.

Although the ONT sequencing accuracy has been largely improved in recent years, its error rate is still higher than Illumina's. Therefore ONT output is commonly analysed with the assistance of the

Illumina dataset. In general, the consistency between the two technologies is comparable when investigating gene expression. Using single-cell sequencing, Byrne et al. (2017) showed Pearson's *r* of 0.84 to 0.92 between gene expression level quantified by ONT and Illumina methods. A study that compared mouse gene expression quantification between Illumina Truseq and ONT direct RNA sequencing revealed Spearman's ρ = 0.77 (Sessegolo et al., 2019). In our study, the comparisons between gene expression quantified from ONT and Illumina Truseq sequencing libraries (Figure 6.2) shows Spearman's ρ > 0.8, indicating that the ONT approach largely replicates the Illumina gene expression. The quantification of transcription of TE families showed a moderate correlation with Spearman's ρ = 0.58 to 0.64 (Figure 6.3). The repetitive nature of TEs may have caused the relatively lower correlation between ONT and Illumina data in TE analysis than that observed in the analysis of gene expression. In fact, it has been reported that most of the difference between ONT- and Illumina-based assemblies for *Saccharomyces cerevisiae* genome resides in the repetitive region (Goodwin et al., 2015). In the study of Goodwin et al. (2015), the ONT-based assembly of the *S. cerevisiae* genome largely filled the gaps in TE-enriched regions, where the resolution was poor in the Illumina-based assembly. However, a correction of the systemic sequencing error on ONT sequencing reads by utilization of short-sequencing reads is required prior to performing *de novo* assembly based on ONT reads (Goodwin et al., 2015). A similar situation might have happened in our ONT cDNA sequencing dataset, in which ONT cDNA reads derived from TEs were capable of retaining intact TE transcript information that was fragmented in our Illumina dataset, and therefore resulted in a lower correlation between the two datasets in TE expression analysis (Figure 6.3) than that in gene analysis (Figure 6.2). Nonetheless, the alignment quality of the ONT cDNA reads on TEs, especially full-length TEs, might tend to be affected by the systemic sequencing error of this technology, thus combined analysis of the ONT and Illumina sequencing datasets might help improving this problem.

In the collection of expressed TE loci, a huge proportion of TEs with at least one ONT read was not found in the expression candidate pool collected from the Illumina dataset (Figure 6.4 A, Figure 6.5 A). The majority of these TE loci uniquely identified in the ONT dataset were found to have only one ONT read with an extremely low breadth of coverage (Figure 6.4 B, Figure 6.5 B). As mentioned above, this might be related to aberrant alignment caused by the error-prone ONT reads. By contrast, TE loci supported by both sequencing methods generally show a higher level of ONT read counts and breadth of coverage. These suggest that the incorporation of both sequencing platforms improves the overall expression range of expressed TEs by largely excluding TE loci with poor read count and coverage across features.

### 6.5.2 Validation of the association between TEs and genes found in the Illumina data

In chapter 3, using Illumina data, the expression candidates show a strong location bias towards the intron of expressed genes. To validate this finding, the expressed TEs in ONT data were categorized in the same way. These TEs revealed a very similar distribution pattern (Figure 6.6) to that of the expression candidate pool (Figure 4.14). As reported in chapter 4, there is a negative correlation between the presence of intragenic TEs and gene expression level in the Illumina data. In that case, genes co-localized with full-length expression candidates and genes having no active TEs but containing inactive ones were less likely to be highly expressed. This phenomenon was reproduced in the independent stress treatment using ONT cDNA sequencing (Figure 6.9 – Figure 6.12). The data in this chapter generates similar results and confirms our initial experimental approach using an independent method of sequencing. Thus the works of previous chapters are validated. Investigation of gene alternative splicing associated with TE

Although the inclusion of TEs in gene transcripts could interfere with translation efficiency from multiple aspects, the presence of TEs in Alt3, Alt5, and ES features didn't show a noticeable difference in the predominant types of translation productivity from the non-TE-related counterparts (Figure 6.13 B). In contrast, the majority of isoforms containing TE-related IR had premature termination codons (PTC), whereas isoforms containing IR that were not related to TEs were predominantly estimated to be productive (PRO) in producing protein (Figure 6.13 B). Further analysis found that most of these premature termination sites appeared within the retained introns, nine of which nested in the TE sequences. These demonstrate that intronic TEs can influence gene expression, in terms of the efficiency to produce protein, by introducing premature stop codon within the unspliced intron. These TEs may serve alternative polyadenylation sites (Kuang et al., 2009; Tsuchiya and Eulgem, 2013) that cause accumulation of unproductive transcripts.

From an epigenetic perspective, it is plausible that TEs retained in introns can act to regulate alternative splicing through epigenetic-related chromatin re-organization. Evidence has accumulated for the co-occurrence of transcription and splicing, i.e. splicing of a gene transcript takes place while the polymerization of this transcript is ongoing. Accurate splicing requires the delicate coordination among RNA Polymerase II (Pol II), various specialized proteins and splicing signals around the spicing sites (Barash et al., 2010; Naftelberg et al., 2015; Reddy et al., 2012). Furthermore, chromatin structure also plays a crucial role in defining exons and directing exon selection in the splicing process (Jabre et al., 2019; Naftelberg et al., 2015; Reddy et al., 2012). The multilayered coordination between histone modification, DNA methylation, and siRNA biogenesis underlies chromatin organization. Histone acetylation is one of the histone modifications that directly affect chromatin-

histone attraction, in which the negatively charged histone acetyl-tail repels the negatively charged DNA, resulting in a more relaxed conformation (Naftelberg et al., 2015). Inhibition of mouse histone deacetylase (HDAC) resulted in alternative splicing of roughly 700 genes, possibly through accumulated histone H4 acetylation and increased processivity of Pol II (Hnilicová et al., 2011). In addition, RdDM is commonly mediated by siRNA and small RNA-associated protein, Argonaute, and leads to deposition of DNA methylation as well as suppressive histone modifications such as H3K9me2 and H3K27me3. These modifications result in a condensed heterochromatic structure that hampers transcription. It has been proposed that heterochromatic introns may act as intragenic roadblocks to transcriptional elongation and affect alternative splicing (Alló et al., 2009). Given that silenced TEs conventionally bare these epigenetic hallmarks, intronic TEs may influence alternative splicing through the epigenetic marks placed upon them. High levels of H3K9me2 on an intronic *COPIA-R7* in *Arabidopsis* was found crucial in preventing the use of alternative polyadenylation site hidden by the TEs, therefore secured the functional transcription of the *Arabidopsis* disease resistance gene *RPP7* (Tsuchiya and Eulgem, 2013). Clonal propagation of oil palm increases the tendency for hypomethylation and the loss of 24-nt siRNA targeting the intronic TE located in *Karma* (Ong-Abdullah et al., 2015). This alteration of 24-nt siRNA further contributed to aberrant exon skipping that introduced premature stop codon to the host homeotic gene *DEFICIENS* and thus resulted in fertile mantles phenotype of oil palm. Taking these results together, it is sensible to speculate that the epigenetic modifications laid on these intronic TEs may act to secure correct splicing. This could mask unwanted signals, including, but not limited to, alternative splicing enhancers and premature stop codon hidden in the TE-containing introns but not necessarily in the TEs. When the epigenetic suppression on these TEs is lifted, the unmasked intron would have higher possibility of causing aberrant splicing that exposes premature stop site in the mRNA. Returning to the findings in 6.4.4, TEs retained in introns do not necessarily carry or provide premature stop codon directly; instead, they may work to leverage splicing regulation coupled to necessary activity of the epigenetic machinery.

### 6.5.3 Identification of competent transcription from autonomous TEs

With the same analysis workflow, several TE families were likely to have autonomous TE loci with nearly full read coverage over TE domains required for transposition. These TE families include Copia-3, Copia-23, which were the two TE families that have shown full read coverage in the Illumina data, as well as Gypsy-V1, LINE7, LINE8 and hAT-7 that were additionally identified as TE families with full read coverage in the ONT data (Figure 6.15 – Figure 6.17). Nonetheless, only Gypsy-V1 and hAT-7 showed that the full coverage was contributed by intact ONT reads instead of an assembly from partially overlapped ONT reads (Figure 6.25, Figure 6.26). These findings suggest that Gypsy-V1 and

hAT-7 have the potential of producing autonomous transcripts with the wound-like treatment. Moreover, it also indicates that the so-called 'TE activation' commonly seen from short-read RNAseq may not necessarily represent transcriptional activation competent for autonomous mobilization of elements.

One might wonder why there was no autonomous Gypsy-V1 and hAT-7 loci identified as the potential origins of the competent transcript in the Illumina data. From the genome browse images of Figure 6.25 and Figure 6.26, although with the presence of ONT reads covering through the necessary domains, the expression level of these TE loci seemed low. The short sequencing reads are not able to retain intact information of a long transcript produced from an autonomous TE locus, especially when the transcriptional activity of this locus is low, while an ONT read is able to represent a single long transcript. The low transcriptional level observed in the ONT data may also reflect bias in the cDNA sequencing due to PCR amplification and preferential order of ONT sequencing (i.e. short reads first).

Given the low level of transcription observed, does the threshold of one ONT read strong enough to represent TE activity? Firstly, these reads were initially selected with the presence of flanking primers, which indicate the polyadenylation of the transcripts as well as completed reverse transcription and strand switching during library preparation. As a result, each of the selected reads generally represents a full-length transcript, accounting for the transcriptional activity. Secondly, the identified autonomous loci were supported by the Illumina data, where these loci were considered as expression candidates. It appears that the presence of one ONT read is sufficient when there is also evidence from short-read based sequencing.

Undeniably, the more the read depth, the more confidence in transcriptional activity. Panda and Slotkin (2020) set five ONT reads as the threshold to annotate an active TE locus. However, they pooled six libraries, which included *Arabidopsis* wild-type and four epigenetically compromised genotypes, before applying this threshold. This generated more than 5 million reads for a model organism with a genome size of ~135 Mb. With a similar N50, it would require alignment depth at least 2.3 times deeper than the current mock or yeast libraries that we have interrogated. The *Arabidopsis* genotypes used in their study included those which were compromised in epigenetic silencing that allows the constant de-repression of TE loci, whereas the grapevine embryogenic callus was initiated from wild-type *P. noir* clone UCD5 and was subjected to temporal or continuous stress treatment. In such a fully active epigenetic silencing environment, the likelihood of high levels of transcription of autonomous elements is low. That being said, providing a stress treatment that can lead to strong transcriptional activation of autonomous TE loci in the wild-type background, more sequencing and alignment depth may be necessary to detect the full-length TE transcription than

normally required to detect gene activation. For this purpose, size selection for cDNA longer than a certain threshold may lower the requirement of sequencing depth. Alternatively, using genetic backgrounds predisposed to TE activation, such as appropriate mutations in key epigenetic regulators, applying chemicals that inhibit proteins of the silencing machinery, or conducting stress treatment more specific to the stress-responsiveness of particular TEs may raise the visibility of autonomous TE transcription.

### 6.5.4 Survey of the stress-related CREs in LTRs of representative TE families

Despite the full-length transcripts of Copia-3 and Copia-23 were undetectable in the ONT data, from the perspective of high sequence identity within the family (see Chapter 4) and overall transcriptional activity, they are still two of the most promising LTR-TE families with mobilization potential.

Copia-3 and Copia-23 have INT domain two to three times longer than Gypsy-V1's INT, which could have increased the risk of being targeted by post-transcriptional silencing and expression-dependent RdDM (Cuerda-Gil and Slotkin, 2016; Panda et al., 2016). By comparison, the canonical LTR sequence of Gypsy-V1 is three to four times larger than that of Copia-3 and Copia-23. This leaves more CREs nested in Gypsy-V1's LTR and fewer CREs in Copia-3 and Copia-23 (Figure 6.25). Together, these might influence their various degree of sensitivity to stress treatments and different visibility of autonomous transcripts in the transcriptome.

The CREs annotated in the LTRs imply types of stress treatments that could be effective for TE activation. However, the multi-layered epigenetic network may only leave a very small and transient window for TE activation as responding to stress. The silencing strength after the transcriptional activation might be even stronger than before (Marí-Ordóñez et al., 2013; Secco et al., 2015). Taken together, custom-designed stress treatment to specific TEs coupled with measures that ease epigenetic suppression of TEs may help to achieve the aim of increasing efficient mobilisation of autonomous TEs.

## 6.6 Conclusions

The contiguous nature of the ultra-long read sequence provided by ONT sequencing is revolutionising. This particularly applies to the detection of transcript isoforms and alternate splicing in transcript data and largely facilitates the study of TE transcriptional activity. In this chapter, combined analysis of the ONT and Illumina dataset can further decrease the size of TE expression candidate pool, in which expressed TE loci identified by both sequencing platforms were generally supported by more breadth of coverage and read count of ONT reads than were the TE loci lack evidence from short-read data.

The ONT cDNA dataset recapitulated the observations from the short-read sequencing data (chapter 3 and chapter 4), including the location bias of expressed TEs and the negative association between the presence of TEs and the expression level of co-localized genes. In addition to that, as revealed from the alternative splicing analysis, intragenic TEs could be involved in gene regulation through participating in alternative splicing. Particularly, intron retention overlapping with TEs tends to associate with exposure of premature termination codons in transcripts. Although it has been reported that intronic long TEs may be spliced out yet interfere with long-range recognition of constitutive splice sites, the alternative splicing analysis in this chapter only captures alternative splicing features directly overlapping with TEs. Therefore the long-range alternative splicing related to splicing long TEs was not examined in this dataset. However, the applied analysis approach provides solid observations regarding the systematic survey of TE-overlapping alternative splicing. Last but not least, the ONT cDNA libraries were able to reveal consecutive transcripts spanning across autonomous TE loci without assembly. It shows that, under the wound-like experimental settings, autonomous Gypsy-V1 and hAT-7 loci were likely to have full transcription across regions necessary for transposition. On the contrary, Copia-3 and Copia-23, the two LTR-TE families that revealed the most promising transcriptional activity in the Illumina data and were considered having experienced the latest mobilization burst in evolutionary time in grapevine, did not show competent full-length transcription in the ONT data. Even each of the aforementioned autonomous Gypsy-V1 and hAT-7 loci demonstrated merely one confident ONT read. A closer examination for stress-related CREs in LTRs of Copia-3, Copia-23 and Gypsy-V1 suggests that wound-like or biotic stress treatments may not be the most efficient way to stimulate significant *de novo* transcription of these TEs. In addition to using more effective stimuli based on the annotated CREs, it's plausible that pharmacological inhibition of the epigenetic machinery may be required to fuel the efficient stress-responsiveness of autonomous TEs.

# Chapter 7

# Analysis of TE transcriptional activity with pharmacological inhibition of histone deacetylase

## 7.1 Overview

Post-translational modifications on histone tails are crucial for the three-dimensional conformation of chromatins, which is tightly associated with the epigenetic landscapes and TE activity. Mutations in genes encoding histone deacetylases (HDACs) or application of HDAC inhibitors (HDACi) have been found to increase the acetylation level on histones and transcriptional activity of TEs. In order to test the ability of HDACi in TE de-repression, grapevine embryogenic callus with wound-like pre-treatment were continuously incubated with the HDACi trichostatin A (TSA) and 4-phenylbutyric acid (4PBA) across a time series. Short-read RNAseq data showed that 4PBA is more effective than TSA in terms of TE activation. Without increasing the total number of potentially expressed TE loci (so-called expression candidates), 4PBA treatment gave rise to a new subset of ~2,500 expression candidates that encompassed a wider range of active TE families that were absent in the mock treatment. Taken together with the changes in location distribution mostly contributed from the 4PBA-specific expression candidates, it seems that 4PBA changes the landscape of TE activity, not by globally broadening transcriptionally permissive regions, but by shifting the spectrum of the permissive transcriptional area within the genome and with respect to TE transcriptional loci. In addition, the transcriptional activity of genes is likely to be a determinant in defining TE permissive regions. This regional activation with 4PBA is also supported by the existing evidence for other studies, which show that 4PBA generally inhibits a subset of HDACs and that most of the HDACs preferentially target active or inducible genes primed by H3K4 methylation rather than deeply silenced genes and subtelomeric regions. The shift in TE permissive area might also echo the gene ontology networks significantly affected by 4PBA. However, the TE activation seen in the wound-like, biotic-stress, and 4PBA treatments did not result in massive production of full-length TE transcripts that are detectable by ONT cDNA sequencing. It is proposed that combined use of pharmacological inhibitors of epigenetic silencing machinery and suitable environmental cues is likely to be required for generating effective transposition bursts in the absence of pre-existing mutations in epigenetic machinery in wild-type backgrounds.

## 7.2 Introduction

Histone modifications play crucial roles in modulating chromatin structure, which is a determinant of the accessibility of underlined genes to the transcriptional machinery. Over 100 distinct histone modifications have been discovered (Zentner and Henikoff, 2013), among which lysine methylation and acetylation are the most described.

As mentioned in chapter 1, methylated histone 3 at lysine 4 (H3K4me) and H3K36me are frequently found in euchromatin and indicative of active and inducible transcription, whereas methylated H3K9 and H3K27 modified by histone methyltransferase are hallmarks frequently associated with silenced TEs in constitutive heterochromatin or repressed genes in euchromatic regions (Berger, 2007; Pfluger and Wagner, 2007). Histone acetylation and deacetylation are catalysed by histone acetyltransferase (HAT) and histone deacetylase (HDAC), respectively. High levels of acetylated histones are a characteristic of chromatins associated with transcribed genomic regions, while low levels of acetylated histones are linked with transcriptionally inactive or deeply silenced areas (McAnena et al., 2017). The majority of predicted autonomous TEs in plant genomes are often associated with heterochromatic regions enriched with H3K9me or H3K27me or showing a paucity of acetylated histones (Shahbazian and Grunstein, 2007; Sigman and Slotkin, 2016).

Since histone methylation and acetylation at histone N-terminal tails are reversible modifications, pharmacological intervention in the corresponding enzyme activity may de-repress TE activity. Several histone deacetylase inhibitors (HDACi) have been discovered, with their target specificity and enzyme activity being reported in animals, plants and have been used in clinical trials for treating human cancers (Bolden et al., 2006; Falkenberg and Johnstone, 2014; Ma et al., 2013). In contrast, the variety of histone methyltransferase inhibitors, as well as the breadth and depth of their functionality, are less well reported (Bissinger et al., 2010; Kubicek et al., 2007). Therefore, this chapter focuses on the impacts of HDAC and HDACi on TE activity.

### 7.2.1 Histone deacetylase

HDAC has been described as an epigenetic eraser, which removes epigenetic marks from DNA or histones after these marks being deposited or recognized by epigenetic writer and reader, respectively (Falkenberg and Johnstone, 2014). Animal and plant HDACs are basically categorized according to their homology to yeast HDACs. In humans, the HDACs homologous to the yeast Rpd3, Hda1 and Sir2 proteins are classified as class I, class II, and class III, respectively (Bolden et al., 2006; Thiagalingam et al., 2003). An additional class IV HDAC exists in human because human HDAC11 shares sequence similarity at its catalytic core with both class I and class II enzymes, yet the overall identity with either class I or II enzymes are low, necessitating its classification as a new class of

enzymes (Bolden et al., 2006; Gao et al., 2002). To date, 18 HDACs have been discovered in *Arabidopsis*, in which the RPD3/HDA1 group (12 HDACs) is the biggest HDAC family, following by the SIR2-like (2 HDACs) and HD2-like groups (Pandey et al., 2002). The HD2 enzymes were firstly discovered in maize (Lusser et al., 1997) and recognized as plant-specific HDACs, which have not been identified in animals (Dangl et al., 2001; Ma et al., 2013). The RPD3/HDA1-like HDACs can be further categorized by the similarity to yeast Rpd3, Hda1 and Hda2 proteins (Alinsug et al., 2009). The RPD3-like HDACs generally show the nucleus localization and ubiquitous expression in various tissues, whereas the HDA1-like HDACs can shuttle between the cytoplasm and the nucleus (Alinsug et al., 2009; Bolden et al., 2006). On the other hand, the SIR2-like HDACs are $NAD^+$-dependent and thus regulated by the AMP-activated protein kinase (AMPK) signalling pathway (Cantó et al., 2009; Imai et al., 2000). The categorization of *Arabidopsis* HDACs is summarized in Table 7.1.

**Table 7.1 HDACs in Arabidopsis.**

This list is sorted according to Alinsug et al. (2009) and Pandey et al. (2002).

| HDAC families | Subfamilies | Enzymes |
|---|---|---|
| RPD3/HDA1 | RPD3 | AtHDA6 |
| | | AtHDA7 |
| | | AtHDA9 |
| | | AtHDA10 |
| | | AtHDA17 |
| | | AtHDA19 |
| | HDA1 | AtHDA5 |
| | | AtHDA8 |
| | | AtHDA14 |
| | | AtHDA15 |
| | | AtHDA18 |
| | HDA2 | AtHDA2 |
| HD2 | | HDT1 (AtHDA2A) |
| | | HDA2 (AtHDA2B) |
| | | AtHDA3 |
| | | AtHDT4 |
| SIR2 | | AtSRT1 |
| | | AtSRT2 |

## 7.2.2 Roles of histone deacetylase in epigenetic silencing

The aberrant expression of HDACs and recruitment of HDACs to oncogenes are widely observed in various human tumour tissues and cancer cells (Falkenberg and Johnstone, 2014; Witt et al., 2009). In leukaemogenesis, HDACs can be recruited by leukaemia-associated fusion proteins, like acute myeloid leukaemia 1 (AML1)-ETO and promyelocytic leukaemia-retinoic acid-related receptor-α (PML1-RARα), and form multi-protein complexes with DNA methyltransferases (DNMTs) to silent genes promoting cell differentiation (Falkenberg and Johnstone, 2014; Minucci and Pelicci, 2006). In

*Drosophila*, the HDAC dRPD3 can be co-immunoprecipitated with the chromatin remodeler ISWI, which mediates chromatin compaction potentially by facilitating the association of linker histone H1 and chromatin (Brehm et al., 2000; Corona et al., 2007). In *Arabidopsis*, the RPD3-like AtHDA6 was found responsible for rRNA gene silencing as knockdown of AtHDA6 resulted in decreased cytosine methylation at the promoters and the replacement of H3K9me2 with H3K4me3 (Earley et al., 2006). There is evidence that shows AtHDA6 directly interacts with DNA methyltransferase (MET) and takes part in TE and heterochromatin silencing (Liu et al., 2012; To et al., 2011b). Moreover, the interaction of AtHDA6 protein and the histone methyltransferase SUVH5 has been proved by in vitro pull-down, co-immunoprecipitation and yeast two-hybrid assays (Yu et al., 2017). The synergetic effects of AtHDA6 and the histone methyltransferases in establishing H3K9me2 and erasing H3K9K14ac were revealed by the quadruple mutant impaired in AtHDA6 and SUVH4/5/6 (Yu et al., 2017). Another RPD3-like HDAC, AtHDA19, was found to form a repressor complex with brassinosteroid (BR)-related transcriptional factor BES1 (Ryu et al., 2014). BES1 has been shown to be able to recruit the histone demethylase ELF6 (Early flowering 6) that demethylate H3K27me3 to H3K27me1 (Antunez-Sanchez et al., 2020; Crevillén et al., 2014; Yu et al., 2008). These together demonstrate that HDACs participate in epigenetic regulation in concert with suppressive DNA methylation and histone methylation and possibly interact with histone demethylase as well.

### 7.2.3 Histone deacetylase inhibitors

Due to the link between poor prognosis of cancer patients and high level of HDACs, several histone deacetylase inhibitors (HDACi) have been discovered and tested. While some of the HDACi can broadly inhibit all classes of HDACs, others exhibit target specificity. Trichostatin A (TSA) is a pan-inhibitor that can effectively suppress the deacetylation activity of all 11 known human HDACs, while a short-chain fatty acid, butyrate, is preferentially selective for all of class I human HDACs (RPD3-like HDACs) and some of the human class II HDACs (Bolden et al., 2006). Trichostatin A and butyrate have been shown to elevate the level of histone acetylation in many plant species, such as maize (Ransom and Walton, 1997), beans (Belyaev et al., 1997), *Medicago sativa* (Waterborg and Kapros, 2002), tobacco (Kurita et al., 2017; Li et al., 2005), and *Arabidopsis* (Chang and Pikaard, 2005; Li et al., 2014; Rangani et al., 2015). Accompanied with histone hyperacetylation, increased susceptibility of total genomic DNA and transferred DNA to DNase I digestion was also observed in HDACi treated maize cell culture (Tiricz et al., 2018). In most of the plant cases, TSA and butyrate were found to promote seed germination or embryogenesis and induce alterations in auxin, cell wall and cell cycle pathways (Nguyen et al., 2020). These studies in plants also show that TSA works at μM level, whereas butyrate works at mM level (Bolden et al., 2006). Although plant seeds, seedlings or cell cultures were affected by TSA or sodium butyrate with various treated period of time, most of the studies showed

considerable effects in 48 hours (Chang and Pikaard, 2005; Hayashi and Takaiwa, 2015; Kurita et al., 2017; Li et al., 2005; Tiricz et al., 2018). Furthermore, the study in alfalfa revealed the turnover rate of histone acetylation as short as 0.5 hours and that the acetylation level was doubled by TSA within 4 to 6 hours of incubation (Waterborg and Kapros, 2002). In tobacco BY-2 cells, 0.5 µM of TSA effectively resulted in histone hyperacetylation in 1 hour of treatment, and displayed gradual increases in H3K9 acetylation level in 3 and 6 hours. Given the advantage of pan-inhibition activity at low concentration level, TSA was not tested in clinical trial until recently (https://clinicaltrials.Gov accessed on 5 July 2020), whereas a chemical derivative of butyrate, phenylbutyrate, has been through phase I and phase II for its application in treating cancer in 2014 (Mottamal et al., 2015), and approved by the Food and Drug Administration as a safe drug for patients with hyperammonemia (Iannitti and Palmieri, 2011; Kusaczuk et al., 2016). Phenylbutyrate (or sodium phenylbutyrate, 4-phenylbutyric acid, 4PBA) is a stable butyrate derivative (Kusaczuk et al., 2015). This compound is equipped with a phenyl group that additionally grants it chaperone-like properties, which was not seen in unmodified sodium butyrate. Therefore, while the cytotoxicity of TSA has been discussed (Alao et al., 2006; Blagosklonny et al., 2002, 2005), 4PBA was proved, in mammals, to display cytoprotective properties in easing endoplasmic reticulum (ER) stress via its chaperoning activity, which is important for protein folding (Kusaczuk et al., 2015). Studies have shown that 4PBA restored ER homeostasis in neuronal cells suffering from ER stress (Kubota et al., 2006; Wiley et al., 2010). 4PBA's roles in suppressing histone deacetylation, anti-cancer activity, ammonia scavenging, and ER homeostasis have been widely investigated in vertebrates (Gore and Carducci, 2000; Kusaczuk et al., 2015). However, its chaperon-like activity is more frequently reported than the deacetylation role in plants suffering from ER stress-induced by abiotic and biotic stimuli (Avin-Wittenberg, 2019; Hayashi and Takaiwa, 2015; Watanabe and Lam, 2008; Yang et al., 2016). In most of these cases, 3 to 12 hours of treatment with 1 to 2 mM of 4PBA treatment is sufficient to evoke ER homeostasis. These together raise the question of whether 4PBA can provoke TE de-repression by acting as HDACi yet exhibit less cytotoxicity.

### 7.2.4 Research workflow

Given the epigenetic role of HDACs, it is hypothesised that chemical inhibition of HDACs may de-repress the transcriptional activity of TE loci in chromatin regions where the HDAC activity is more dominant than HAT. To test this aim, two types of HDACi, TSA and 4PBA, were applied to the grapevine embryogenic callus pre-treated wound-like procedure (see chapter 3). As shown in Figure 7.1, this would involve a systematic investigation of TE activity in multiple perspectives with the analysis pipeline established in chapter 3 and chapter 4. In addition, using the same approach

described in chapter 6, the data from ONT cDNA sequencing would cooperate into the analysis with Illumina RNAseq data to detect full-length TE transcripts derived from autonomous TE loci.

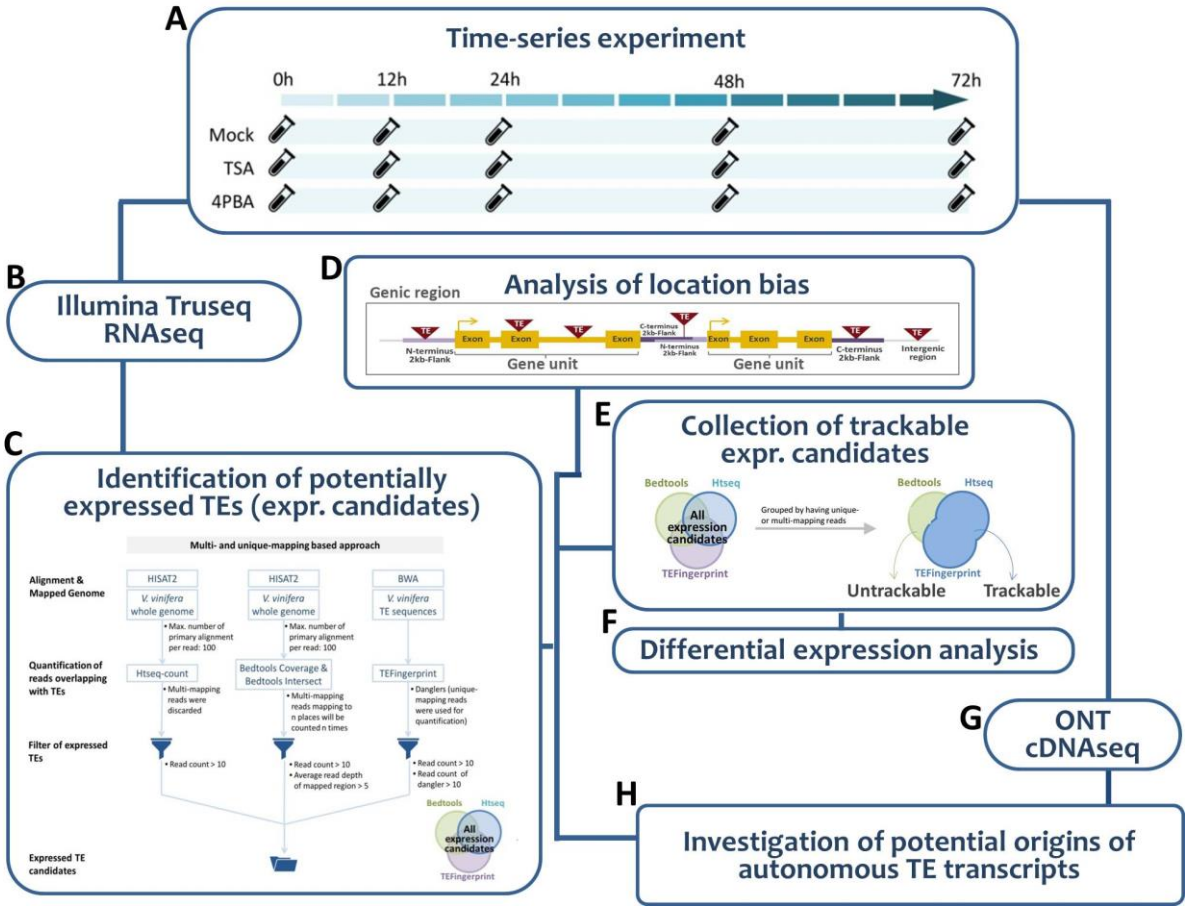



**Figure 7.1 Research workflow**

**(A)** Experimental design of time-series HDACi treatment. The embryogenic calli were harvested at the indicated time points, and the RNA samples were collected. **(B)** The RNA samples were prepared for Illumina Truseq sequencing. **(C-F)** The identification of transcriptionally active TE loci from the sequenced libraries was conducted based on the analysis pipeline (C) established in chapter 3, following by the characterization of active TE loci by location (D), integrity, presence/absence of unique mapping reads (E) etc. TE loci having unique-mapping reads (a.k.a. trackable expression candidates) were further investigated for transcriptional dynamics across treatments over time (F). **(G-H)** The RNA samples were also prepared for ONT cDNA sequencing (G). The resulting data was co-analysed with the Illumina dataset to identified TE loci that may contribute full-length transcripts necessary for autonomous mobilization.

## 7.3 Methods

### 7.3.1 Stress treatment

The embryogenic callus was established as described in chapter 2. The treatment series consisted of a mock treatment that included a vigorous mechanical shaking that would stimulate a wound-like response in the callus (also see section 2.3.1), a treatment combined mock and 0.5 µM of TSA (Sigma, stock dissolved in DMSO), and a treatment combined mock and 2mM of 4PBA (Sigma, stock dissolved in sterile water). The mock plus TSA treatment was hereafter simply denoted as TSA treatment, while the mock plus 4PBA treatment was denoted as 4PBA treatment. Callus of the mock treatment was harvested after 12, 24, 48 and 72 hours of manipulation, recovering on C1P plate. The mock plus drug-treated callus involved 12, 24, 48, and 72 hours of continuous incubation with the HDACi before harvesting (Figure 7.1 A). Each time point included three technical replicates. Given the scale of the experiment, the mock, TSA and 4PBA treatments were conducted on different but consecutive days. To establish the reference points comparable across treatments, untreated callus were collected and serve as time zero (T=0) on the day of each treatment.

### 7.3.2 RNA extraction, Illumina Truseq and ONT cDNA library preparation, and sequencing

Total RNA was extracted as described in chapter 6. The Illumina RNAseq library was prepared following the protocol of the KAPA mRNA HyperPrep kit (Roche) and sequenced on the Illumina HiSeq X-TEN platform. The long read cDNA library was prepared and sequenced using MinION sequencing by a fellow lab member, Mark Stoop, using methods as described in chapter 6. The Illumina sequencing data were processed and analysed following the protocols in chapter 2 to chapter 4, while the computational analysis of the ONT cDNA sequencing data was conducted in the same manner as described in chapter 6.

### 7.3.3 Selection of grapevine genes potentially involving in epigenetic silencing

A list of genes (Appendix C.8) taking part in *Arabidopsis* post-transcriptional gene silencing (PTGS) and RNA-dependent DNA methylation (RdDM) pathways were searched against the gene function annotation file established by (Díaz-Riquelme et al., 2016). Matched grapevine genes were further selected if the E-value ≤ 1.00E-20 and the identity between the grapevine gene and *Arabidopsis* gene were higher than 50%. The resulting gene candidates were shown in Appendix C.9.

## 7.4 Results

### 7.4.1 Alignment statistics

The Illumina Truseq RNA sequencing generated roughly 60 to 100 million reads per library, except for a library of 4PBA treatment at the 24-hour harvest time-point (Table 7.2). The alignment rate for mapped reads to the grapevine reference genome was found to be between 80% and 90% relative to total sequenced reads. This includes 1% to 2% of total sequenced reads derived from annotated TE loci. The library with insufficient sequencing depth was then excluded from the following analysis.

**Table 7.2 Mapping statistics for Illumina Truseq RNA-seq.**

| Sequenced libraries | | | Sequenced reads | | Adaptor removal | | Filter tRNA/rRNA | | Mapped reads | | TE-mapped reads | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Cond. | Time | Rep. | | | | | | | | | | |
| Mock | 00 h | a | 81,229,854 | 100% | 79,512,914 | 97.89% | 78,959,704 | 97.21% | 72,482,258 | 89.23% | 1,375,113 | 1.69% |
| | | b | 91,969,354 | 100% | 89,870,130 | 97.72% | 89,408,684 | 97.22% | 82,456,078 | 89.66% | 1,552,110 | 1.69% |
| | | c | 84,195,530 | 100% | 81,636,994 | 96.96% | 80,835,738 | 96.01% | 73,779,378 | 87.63% | 1,429,602 | 1.70% |
| | 12 h | a | 83,617,820 | 100% | 81,880,916 | 97.92% | 81,131,656 | 97.03% | 74,538,745 | 89.14% | 1,233,730 | 1.48% |
| | | b | 96,579,030 | 100% | 94,479,322 | 97.83% | 93,552,170 | 96.87% | 85,778,197 | 88.82% | 1,421,424 | 1.47% |
| | | c | 89,382,446 | 100% | 87,159,506 | 97.51% | 86,200,216 | 96.44% | 78,651,489 | 87.99% | 1,264,915 | 1.42% |
| | 24 h | a | 93,779,958 | 100% | 91,264,406 | 97.32% | 90,224,664 | 96.21% | 82,705,941 | 88.19% | 1,316,232 | 1.40% |
| | | b | 89,723,020 | 100% | 87,510,586 | 97.53% | 86,411,282 | 96.31% | 78,522,854 | 87.52% | 1,178,247 | 1.31% |
| | | c | 88,422,750 | 100% | 86,050,464 | 97.32% | 84,983,080 | 96.11% | 77,865,880 | 88.06% | 1,222,704 | 1.38% |
| | 48 h | a | 92,956,592 | 100% | 90,907,718 | 97.80% | 89,768,448 | 96.57% | 81,858,686 | 88.06% | 1,338,932 | 1.44% |
| | | b | 89,910,956 | 100% | 87,512,294 | 97.33% | 86,743,982 | 96.48% | 79,658,641 | 88.60% | 1,253,113 | 1.39% |
| | | c | 92,217,474 | 100% | 89,979,854 | 97.57% | 88,212,404 | 95.66% | 78,843,831 | 85.50% | 1,241,230 | 1.35% |
| | 72 h | a | 76,144,302 | 100% | 74,508,002 | 97.85% | 73,968,024 | 97.14% | 68,403,373 | 89.83% | 936,317 | 1.23% |
| | | b | 93,641,504 | 100% | 91,198,384 | 97.39% | 90,488,836 | 96.63% | 83,357,204 | 89.02% | 1,226,767 | 1.31% |
| | | c | 85,835,426 | 100% | 83,960,354 | 97.82% | 82,885,236 | 96.56% | 75,376,953 | 87.82% | 1,087,475 | 1.27% |
| TSA | 00 h | a | 69,486,784 | 100% | 67,428,738 | 97.04% | 66,440,670 | 95.62% | 59,380,357 | 85.46% | 1,086,522 | 1.56% |
| | | b | 68,230,630 | 100% | 66,716,958 | 97.78% | 66,066,766 | 96.83% | 60,149,247 | 88.16% | 1,028,377 | 1.51% |
| | | c | 82,295,182 | 100% | 80,128,448 | 97.37% | 77,372,238 | 94.02% | 65,659,662 | 79.79% | 1,238,776 | 1.51% |
| | 12 h | a | 56,267,430 | 100% | 54,861,354 | 97.50% | 54,302,654 | 96.51% | 49,509,713 | 87.99% | 693,740 | 1.23% |
| | | b | 74,272,552 | 100% | 72,668,594 | 97.84% | 71,809,006 | 96.68% | 65,599,399 | 88.32% | 871,074 | 1.17% |
| | | c | 79,812,706 | 100% | 76,850,770 | 96.29% | 75,708,818 | 94.86% | 68,664,737 | 86.03% | 1,171,744 | 1.47% |
| | 24 h | a | 85,995,486 | 100% | 83,917,576 | 97.58% | 83,285,056 | 96.85% | 77,123,508 | 89.68% | 1,121,648 | 1.30% |
| | | b | 96,281,556 | 100% | 94,331,874 | 97.98% | 93,714,142 | 97.33% | 87,030,326 | 90.39% | 1,160,230 | 1.21% |
| | | c | 72,736,292 | 100% | 70,879,274 | 97.45% | 68,579,180 | 94.28% | 59,486,258 | 81.78% | 1,253,661 | 1.72% |
| | 48 h | a | 76,216,568 | 100% | 74,347,170 | 97.55% | 72,389,596 | 94.98% | 63,924,113 | 83.87% | 1,228,032 | 1.61% |
| | | b | 92,317,686 | 100% | 90,162,804 | 97.67% | 89,503,566 | 96.95% | 82,763,233 | 89.65% | 1,218,441 | 1.32% |
| | | c | 76,799,770 | 100% | 75,159,662 | 97.86% | 73,434,794 | 95.62% | 65,414,044 | 85.17% | 1,296,367 | 1.69% |
| | 72 h | a | 77,208,128 | 100% | 75,320,054 | 97.55% | 73,421,460 | 95.10% | 64,697,026 | 83.80% | 1,136,867 | 1.47% |
| | | b | 87,765,954 | 100% | 85,593,758 | 97.53% | 85,069,840 | 96.93% | 78,600,940 | 89.56% | 1,202,114 | 1.37% |
| | | c | 62,362,958 | 100% | 60,793,450 | 97.48% | 59,028,862 | 94.65% | 51,589,784 | 82.73% | 1,087,915 | 1.74% |
| 4PBA | 00 h | a | 80,837,318 | 100% | 79,176,496 | 97.95% | 78,695,236 | 97.35% | 72,732,620 | 89.97% | 1,084,891 | 1.34% |
| | | b | 76,587,984 | 100% | 74,753,234 | 97.60% | 72,042,442 | 94.06% | 61,372,780 | 80.13% | 1,620,994 | 2.12% |
| | | c | 83,862,698 | 100% | 81,935,526 | 97.70% | 79,510,736 | 94.81% | 68,742,190 | 81.97% | 1,659,564 | 1.98% |
| | 12 h | a | 77,731,660 | 100% | 75,361,990 | 96.95% | 73,855,136 | 95.01% | 66,020,632 | 84.93% | 857,158 | 1.10% |
| | | b | 64,063,456 | 100% | 62,407,198 | 97.41% | 61,502,836 | 96.00% | 55,627,514 | 86.83% | 737,126 | 1.15% |
| | | c | 93,006,858 | 100% | 90,579,376 | 97.39% | 89,673,000 | 96.42% | 81,948,298 | 88.11% | 1,096,401 | 1.18% |
| | 24 h | a | 1,211,838 | 100% | 1,185,594 | 97.83% | 1,169,724 | 96.52% | 1,054,292 | 87.00% | 12,374 | 1.02% |
| | | b | 94,129,484 | 100% | 91,586,066 | 97.30% | 90,467,174 | 96.11% | 82,449,943 | 87.59% | 1,055,678 | 1.12% |
| | | c | 88,672,312 | 100% | 86,558,920 | 97.62% | 85,466,622 | 96.38% | 77,792,279 | 87.73% | 947,230 | 1.07% |
| | 48 h | a | 110,060,790 | 100% | 107,695,818 | 97.85% | 105,793,840 | 96.12% | 95,513,893 | 86.78% | 1,212,868 | 1.10% |
| | | b | 98,810,976 | 100% | 96,754,498 | 97.92% | 95,112,630 | 96.26% | 85,889,210 | 86.92% | 1,058,275 | 1.07% |
| | | c | 103,241,370 | 100% | 100,898,130 | 97.73% | 99,554,420 | 96.43% | 90,634,781 | 87.79% | 1,089,479 | 1.06% |
| | 72 h | a | 96,645,450 | 100% | 94,554,390 | 97.84% | 92,610,250 | 95.82% | 82,993,680 | 85.87% | 1,027,943 | 1.06% |
| | | b | 100,336,512 | 100% | 98,233,084 | 97.90% | 96,536,710 | 96.21% | 87,310,326 | 87.02% | 1,067,590 | 1.06% |
| | | C | 98,888,372 | 100% | 96,763,000 | 97.85% | 95,432,936 | 96.51% | 86,976,684 | 87.95% | 1,065,732 | 1.08% |

### 7.4.2 Identification of expression candidates

Prior to any treatment, there were 6,363 individual TE loci potentially expressed in the control set (T=0) of embryogenic callus (Figure 7.2 A). With the mock treatment, the number of expression candidates increased by about 1000, reaching the 7,248 TE loci. Within the experimental time period, there were 6,808 and 6,638 expression candidates identified in the presence of TSA and 4PBA, respectively. These expression candidates were represented by 165, 175, 176, and 201 TE families in T=0, mock, TSA, and 4PBA treatments, respectively. Although the total number of 4PBA expression candidates was less than that of mock and TSA treatments, the variety of expressed TE families was increased in 4PBA treatment compared to other experimental conditions.

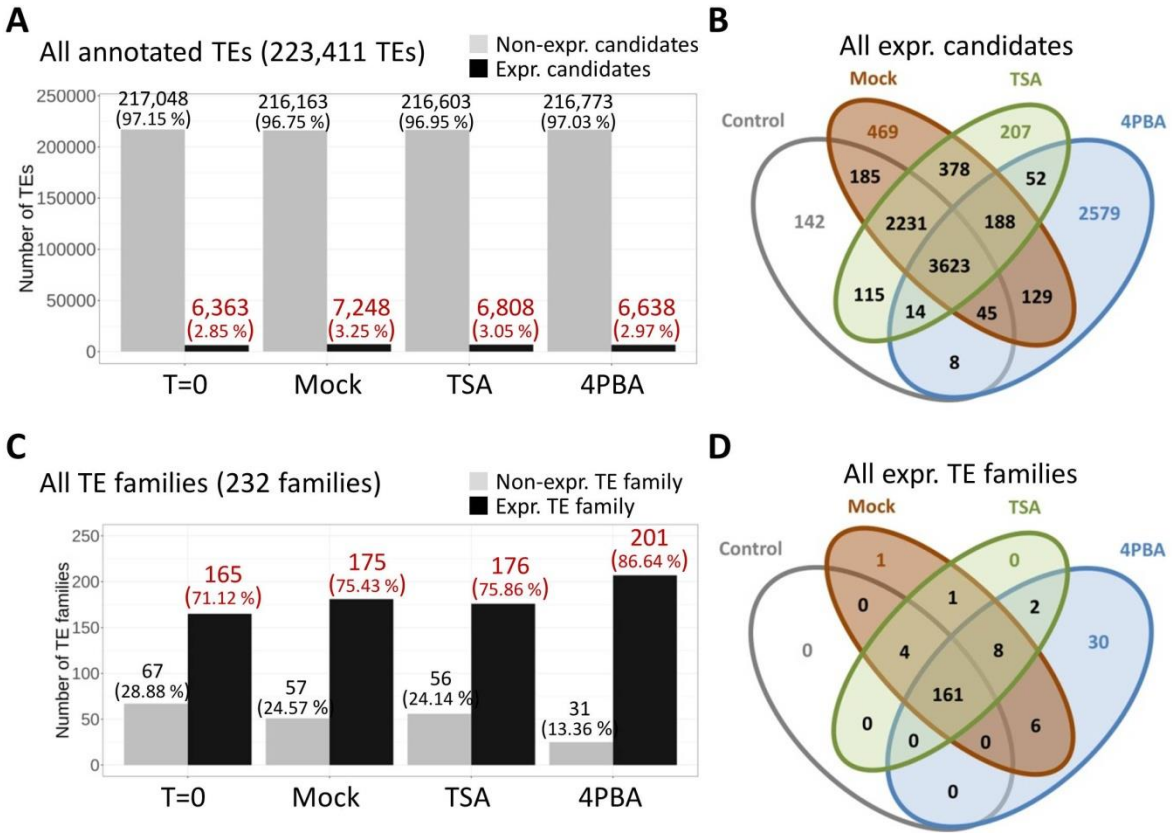



**Figure 7.2 Expression candidates and TE families**

**(A)** For each experimental condition, TE loci passing the analysis pipeline described in chapter 3 were denoted as expression candidates (expr. candidates), whereas the rest would be non-expressed (non-expr.) candidates. For each experimental condition, the numbers of TE loci in these two categories and their percentages relative to all annotated TEs were indicated. **(B)** Comparison of the four sets of expression candidates from all experimental conditions. **(C)** TE families containing expression candidates were considered as expressed TE family (expr. TE family), and the rest was denoted as non-expressed TE family. The number of TE families in each category and their percentage to the total 232 families were indicated. **(D)** Comparison of the four sets of expressed TE families from all experimental conditions.

Given that there was little evidence of HDACi altering the global number of loci becoming transcriptionally active, a cross-comparison of these four sets of expression candidates was carried out and illustrated as a Venn diagram to determine whether any differences in TE loci becoming active may have occurred. The key findings of this analysis are as follows:

(1) Six thousand and eighty-four of the 6,363 expression candidates in T=0 were also identified as expression candidates in mock treatment. Since there are a total of 7,248 expression candidates in mock treatment, there are 1,164 expression candidates in mock not present in the expression candidates pool in T=0.

(2) Comparison of the expression candidate pools of mock and TSA treatments shows that 6,420 TE loci were identified as expression candidates in both treatments. These loci comprised 94% of the expression candidate pool of TSA treatment, suggesting that TSA has little impact in changing the pool of expression candidates compared with mock treatment.

(3) Comparison of the expression candidate pools of mock and 4PBA treatments shows that only 3,985 TE loci were identified as expression candidates in both treatments. There are 3,263 of the 7,248 expression candidates of mock treatment not included in the expression candidate pool of 4PBA treatment, whereas 2,631 of the 6,638 expression candidates of 4PBA treatment were not included in the collection of expression candidates in mock treatment. This analysis revealed differences in total 5,894 TE loci between mock and 4PBA treatments.

When comparing the pools of expressed TE families presented by the expression candidates of different experimental conditions, 165 expressed TE families in T=0 were all found to be presented in the collection of expressed TE families in mock treatments. There were other 10 expressed TE families found in mock treatments but absent in the pool of expressed TE families in T=0. In TSA treatment, 174 of the 176 expressed TE families were also present in the collection of expressed TE families in mock treatment. The small difference in the collection of expressed TE families between mock and TSA treatments is concordant with the previous observation from the comparison of individual expression candidates between mock and TSA treatments. By contrast, there are 32 of the 201 expressed TE families in 4PBA treatments not present in the pool of expressed TE families in mock treatment.

The observations in this section suggest that 4PBA treatment had a stronger impact on shifting transcriptional permissive landscape with respect to expressed TE loci than TSA treatment, and 4PBA treatment is more effective in increasing the variety of expressed TE families than TSA treatment.

### 7.4.3 Location bias of expression candidates

To understand whether different sets of expression candidates show similar distribution patterns in relation to genes, with one or another experimental condition of this chapter, the location of the expression candidates was analysed. The expression candidates of T=0 and mock were largely present in the intron of expressed genes (Figure 7.3 A-C), in concordance with the findings in chapter 3. A similar distribution pattern of expression candidates was revealed in TSA treatment (Figure 7.3 D). There is no significant difference in the distribution of expression candidates in relation to genes among T=0, mock and TSA treatments (Figure 7.4 A-C). However, the proportion of intergenic expression candidates relative to total expression candidates in 4PBA treatment was significantly 7% higher than that in mock treatment (Figure 7.3 E, Figure 7.4 A). In addition, the proportion of expression candidates within flanking region of genes relative to total genic expression candidates in 4PBA treatment was 12% higher than that in mock treatments (Figure 7.3 E, Figure 7.4 B). Compared with mock treatment, the 4PBA treatment also significantly elevated the proportions of expression candidates co-localized with inactive genes, full-length expression candidates, as well as untrackable expression candidates by 5%, 3%, and 5%, respectively (Figure 7.4 C-E). Without the genome-wide TE activation, though, these observations suggest that 4PBA treatment resulted in a shift of transcriptionally permissive area toward intergenic and flanking regions.

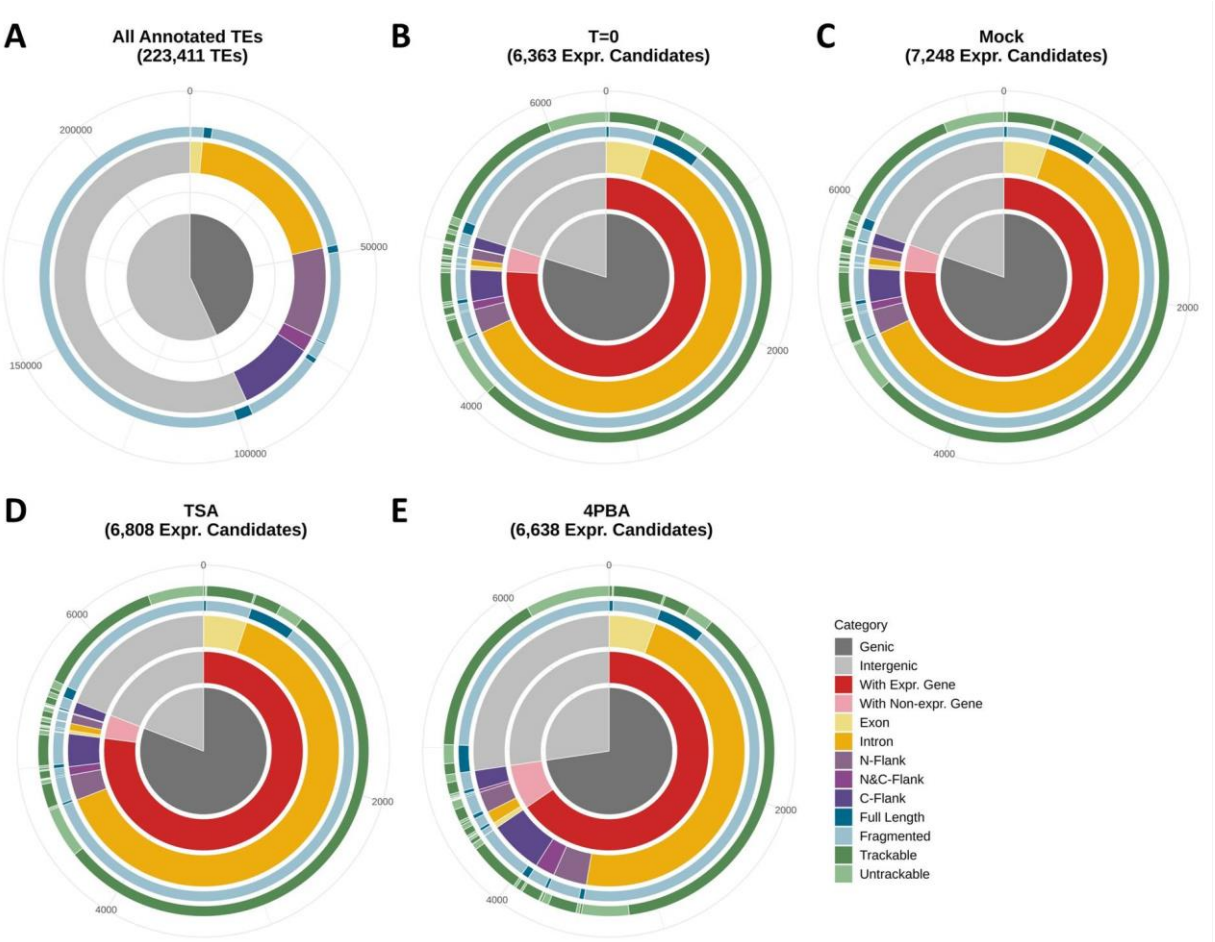



**Figure 7.3 Hierarchical classifications of expression candidates by location, integrity, and distinctness.**

**(A)** All annotated TEs were categorized hierarchically by region (centre), location (internal layer) and integrity (outer-most layer). **(B-E)** Expression candidates of each treatment were categorized in the order of region (centre), the transcriptional activity of co-localized genes (2nd layer), location (3rd layer), integrity (4th layer), and the presence/absence of unique-mapping reads (outer-most layer).

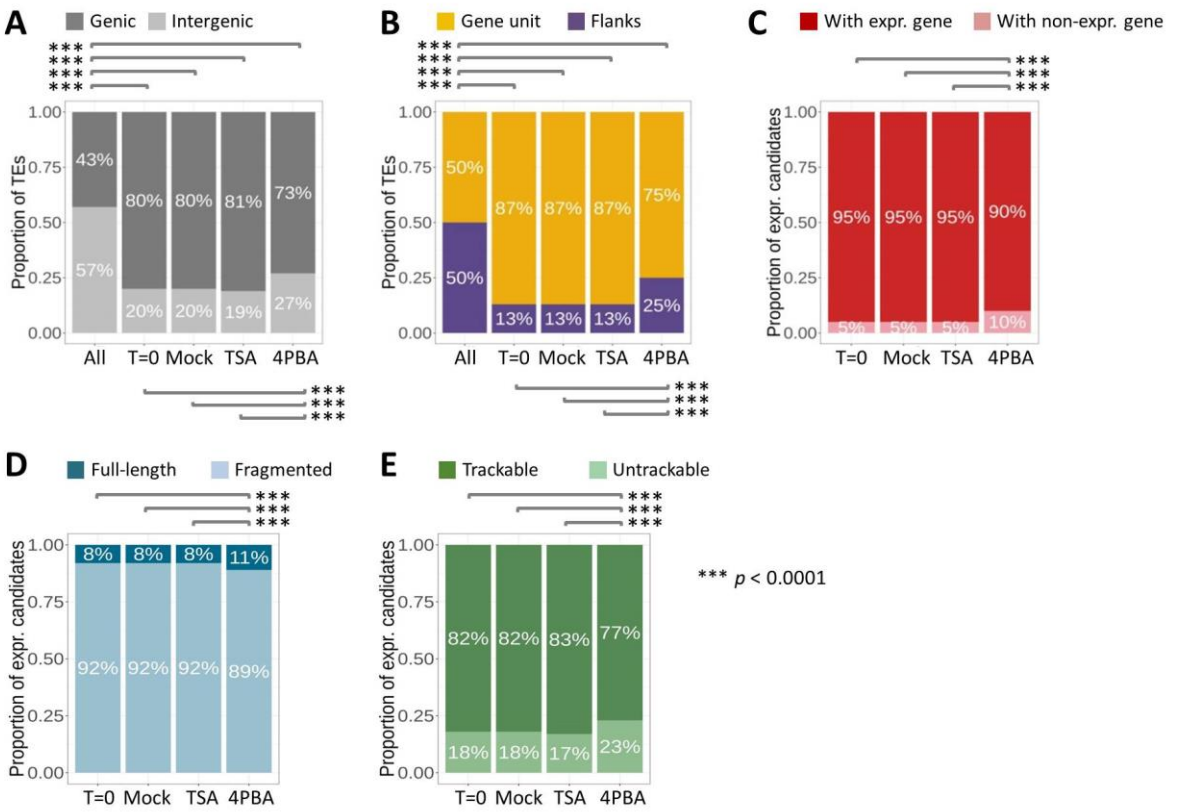



**Figure 7.4 Characteristics of expression candidates in terms of location, integrity and distinctness.**

**(A)** Categorization of annotated TEs and expression candidates by genic/intergenic regions. **(B)** Categorization of annotated genic TEs and genic expression candidates by location relative to genes. **(C)** Classification of genic expression candidates by the transcriptional activity of co-localized genes and statistical comparison between the expected and observed values. **(D-E)** Categorization of all expression candidates by integrity (D) and distinctness (E). The goodness of fit test was performed pair-wisely. All the comparisons reached $p < 0.01$ were labelled. Levels of statistical significance were as indicated.

Previously, the comparison of expression candidate pool between mock and 4PBA treatments revealed that 3,263 expression candidate loci were present in the mock collection but absent in the 4PBA collection of expression candidate (we termed 'mock-unique' expression candidates), whereas 2,653 expression candidate loci were absent in the mock collection but present in the 4PBA expression candidate pool (we termed '4PBA-unique' expression candidates; Figure 7.2 C, Figure 7.5 A). These resulted in 3,985 expression candidates present in both mock and 4PBA treatments (termed 'shared' expression candidates). The significant difference in the location distribution of expression candidates between mock and 4PBA treatments might be either of the two scenarios described below:

(1) It is due to the absence of 3,263 'mock-unique' expression candidate from the 4PBA treatment. In this case, it is expected to observe a significant location distribution difference of 'mock-unique' expression candidates versus the other two ('shared' and '4PBA-unique'),

whereas no significant difference exists between the ‘shared’ and ‘4PBA-unique’ expression candidates.

(2) It is due to the presence of 2,653 ‘4PBA-unique’ expression candidates in 4PBA treatment. If this is the case, it is expected to observe a significant location distribution difference of ‘4PBA-unique’ expression candidates versus the other two (‘shared’ and ‘mock-unique’), whereas no significant difference exists between the ‘shared’ and ‘4PBA-unique’ expression candidates.

To examine these two assumptions, the location distribution of the expression candidates sub-grouped into ‘mock-unique’, ‘shared’, and ‘4PBA-unique’ were analysed and compared with each other (Figure 7.5 – Figure 7.7). The distribution of these three sets of expression candidates revealed that the categorization patterns between the ‘mock-unique’ and the ‘shared’ subsets were similar (Figure 7.5 B, C), whereas the ‘4PBA-unique’ subset showed a considerable increase in the fraction of intergenic expression candidates and those within 2kb distance to genes (i.e. flanking regions) versus all candidates of this subset (Figure 7.5 D). Therefore, it is the presence of ‘4PBA-unique’ expression candidates that contributed to the distribution difference of all expression candidates between mock and 4PBA treatments (Figure 7.3 and Figure 7.4). The statistical test showed that the intergenic proportion relative to all candidates of the ‘4PBA-unique’ subset was significantly higher than the other two (Figure 7.6 A). Although the ‘shared’ subset showed an elevation in flanking proportion compared to the ‘mock-unique’ subset, it is significantly surpassed by the three-fold increase in the ‘4PBA-unique’ subset (Figure 7.6 B). In addition, compared with expression candidates co-localized with expressed genes, the proportion of candidates co-localized with inactive genes in the ‘4PBA-unique’ subset was three to four-fold higher than that of the other two subsets (Figure 7.6 C). These findings mean that although some of the TE loci that were active in mock treatment lost transcriptional activity in the presence of 4PBA, this loss was proportional to the original location distribution seen in the mock samples. In fact, some of the TEs in the intergenic regions (>2kb from genes) and flanking regions (≤2kb from genes) of annotated genes, especially expressed genes, acquired the transcriptional activity in the 4PBA treatment. These observations reinforce the suggestion that, given the time period of our experiment, 4PBA might change the landscape of TE activity, not by globally broadening transcriptionally permissive regions but by shifting the spectrum of permissive area.

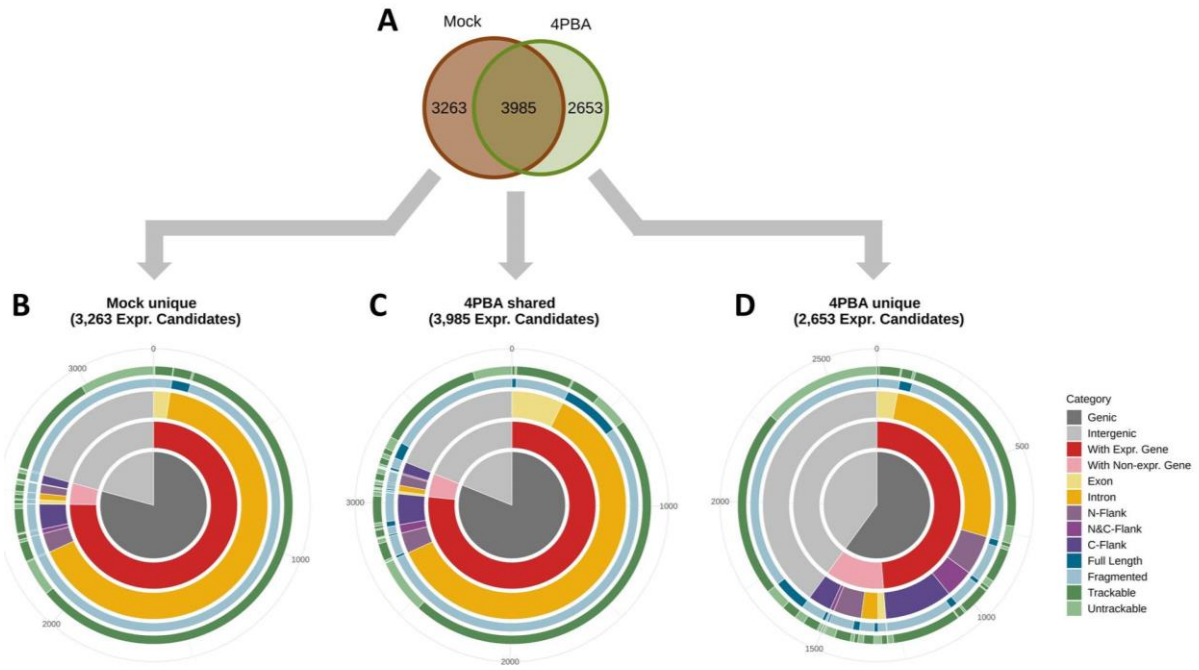



**Figure 7.5 Location biases of expression candidates shared by and unique to mock and 4PBA treatments**

**(A)** Comparison of the identities of mock and 4PBA expression candidates. These candidates were then binned into three groups: mock-unique, shared, and 4PBA-unique. **(B-D)** Hierarchical classifications of expression candidates by location, integrity, and distinctness. These TE loci were categorized as the same as those shown in Figure 7.3.

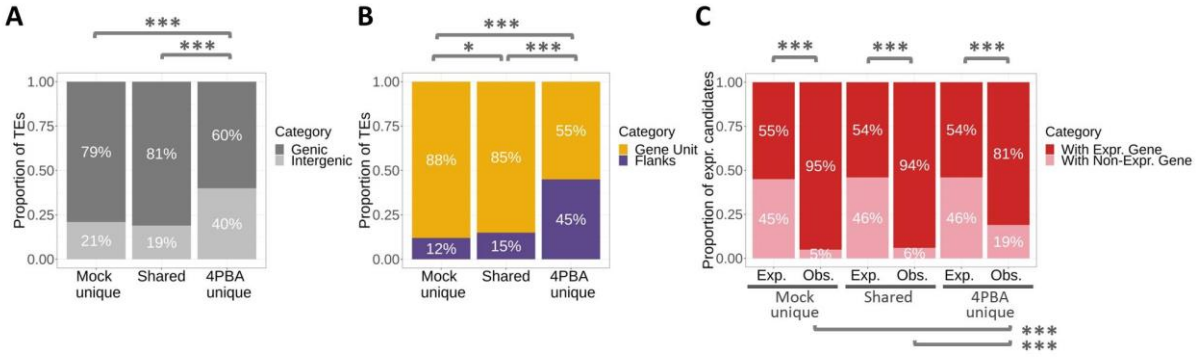



**Figure 7.6 Characteristics of the expression candidates unique to mock, unique to 4PBA and shared by both.**

**(A)** Categorization of annotated TEs and expression candidates by genic/intergenic regions. **(B)** Categorization of annotated genic TEs and genic expression candidates by location relative to genes. **(C)** Classification of genic expression candidates by the transcriptional activity of co-localized genes and statistical comparison between the expected and observed values. The goodness of fit test was performed pair-wise. All the comparisons reached $p < 0.01$ were labelled. Levels of statistical significance were as indicated. Exp., expected; Obs., observed.

It has been reported that various histone deacetylase can target specific genomic regions, and the increased acetylation of histone H3 and H4 is associated with promoters of transcriptionally induced genes (Turner, 2000). Provided that over half of the '4PBA-unique' expression candidates were in the genic region (Figure 7.5 D), it is likely that the gene's transcriptional activity is still a crucial factor defining the permissive area, even for the intergenic expression candidates. In the presence of 4PBA, TE loci located within the area of relaxed chromatin status granted for gene expression might have a higher chance to be expressed than TE loci distal to these areas (i.e. distal to expressed genes). To validate this, the distances from intergenic expression candidates to the closest annotated genes and the closest active genes in 4PBA were examined. If the aforementioned assumption is the case, it is expected that the distance of 'mock-unique' intergenic expression candidates to closest expressed genes in 4PBA treatment is significantly longer than that of 'shared' and '4PBA-unique' intergenic candidates. The comparisons among the 'mock-unique', 'shared', and '4PBA-unique' expression candidates show that the 'mock-unique' intergenic candidates were significantly more distant from the nearest annotated genes than 'shared' and '4PBA-unique' intergenic candidates (Figure 7.7 A, B). The 'mock-unique' intergenic expression candidates were also more distant from the genes active in 4PBA treatment than 'shared' and '4PBA-unique' intergenic candidates (Figure 7.7 C, D). In other words, the expression candidates of 4PBA treatment (including the 'shared' and '4PBA-unique' candidates) were TE loci that are closer on average to transcriptionally active genes, whereas the 'mock-unique' expression candidates absent from the candidate pool of 4PBA treatment were more distal to genes, hence they are less likely to be activated in the presence of 4PBA.

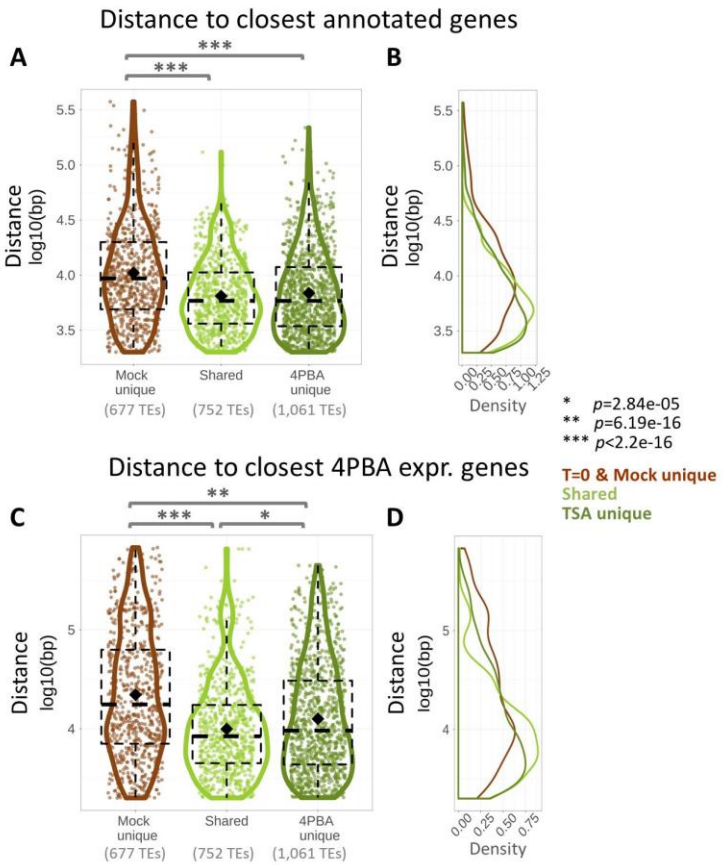



**Figure 7.7 Distance of the mock and 4PBA expression candidates to the closest genes**

The mock-unique, 4PBA-unique, and shared expression candidates were examined for **(A-B)** their distance to the closest annotated genes and **(C-D)** the nearest genes expressed in 4PBA treatment. The distance was logarithmically transformed. Fisher's t-test was performed, and the comparisons with a p-value lower than 0.01 were indicated.

### 7.4.4 Expression patterns of differentially expressed TEs

In addition to the presence and absence of different expression candidates upon various treatments, the responding dynamics of expression candidates in terms of expression patterns is of importance. To explore this concept, expression candidates having unique-mapping reads were gathered for differential expression analysis. Among 5,977 trackable loci in mock treatment, only 32 were differentially expressed TEs (DETEs) in the comparison of mock treatment versus T=0 (Figure 7.8). In the comparison of TSA treatment versus mock and T=0, only 25 of the 5,652 trackable candidates of TSA treatment were identified as DETEs (Figure 7.8). On the contrary, in the 4PBA treatment, there 2,899 trackable expression candidates were identified as DETEs, of which 2,861were unique to 4PBA treatments (Figure 7.8). Note that these DETEs may display various expression patterns, and the shared DETEs may respond to different treatments in distinct ways.

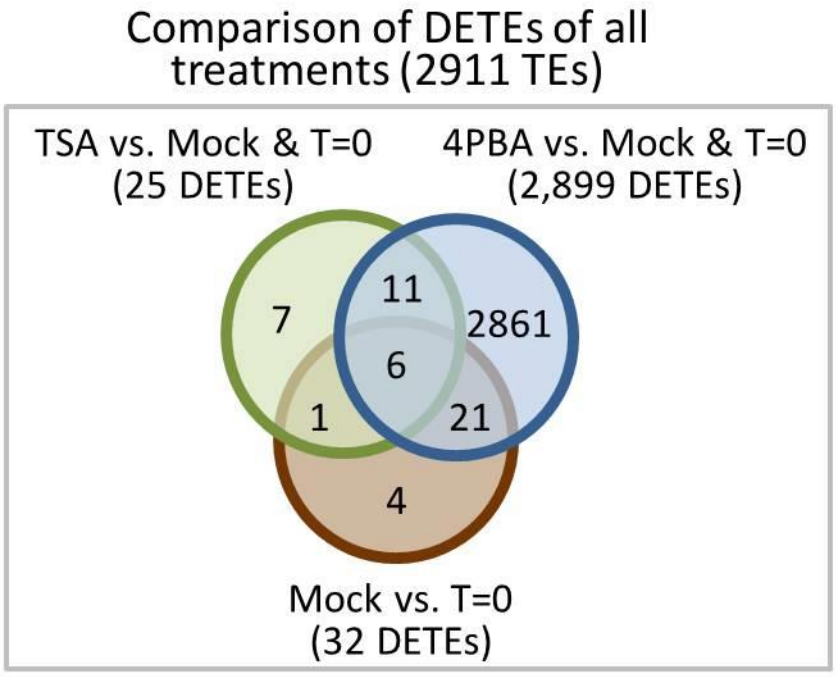



**Figure 7.8 Comparison of differentially expressed TEs of all treatments**

Differentially expressed TEs (DETEs) were identified from the three statistical tests: mock vs T=0, TSA vs mock and T=0, as well as 4PBA vs mock and T=0. The identities of these three sets of DETEs were compared. The numbers of DETEs shared between multiple conditions and those unique to specific treatments were as indicated.

The heatmap and hierarchical clustering analysis show that most of the mock DETEs were once activated 72 hours post the mock treatment. These include 11 DETEs up-regulated across 72 hours of post-treatment period and 17 TE loci that were activated in 24 hours of post-treatment but then returned back to the ground state seen at T=0 (the up-back pattern; Figure 7.9 A-B). In 72 hours of the continuous presence of TSA, 22 of the 25 DETEs were down-regulated compared to mock (Figure 7.9 C, D). By contrast, within 72 hours of 4PBA incubation, 2,059 of the 2,899 4PBA DETEs were up-regulated compared with mock. Among these 2,059 up-regulated DETEs in 4PBA treatment, 1,857 DETEs remained at an elevated level across the time course, while the rest 202 DETEs were activated in 12 to 48 hours of continuous incubation with 4PBA but then returned back to the status resembling T=0 (Figure 7.9 E, F).

As previously shown in Figure 7.5 A, there were 3,985 described as ‘shared’ expression candidates (section 7.4.3) for that these loci were present in the expression candidate pools of both mock and 4PBA treatments, while there were 3,263 and 2,653 expression candidates were termed ‘mock-unique’ and ‘4PBA-unique’ expression candidates respectively (section 7.4.3). In each of these three subsets (‘mock-unique’, ‘shared’, and ‘4PBA-unique’ expression candidates), some of the expression candidates were identified as DETEs in the expression-pattern comparison of 4PBA versus mock

treatments (Appendix C.10 Figure C.15 B-D). It is sensible to speculate that DETEs from the 'mock-unique' subset were down-regulated in the presence of 4PBA and that DETEs from the '4PBA-unique' subset were up-regulated in 4PBA treatment compared with mock. However, it was uncertain how DETEs from the 'shared' subset of expression candidates behaved in the two conditions. Among expression candidates of 'mock-unique', 'shared', and '4PBA-unique' subsets, 560, 985, and 1,354 were identified as DETEs of 4PBA treatment, respectively (note: these DETEs contributed to the total 2,899 DETEs of 4PBA treatments; Appendix C.10 Figure C.15 A-D). It met with the expectations that the DETEs from the 'mock-unique' subset were all suppressed in 4PBA treatment (Appendix C.10 Figure C.15 E) and that the DETEs from the '4PBA-unique' pool were all stimulated (Appendix C.10 Figure C.15 G). For the 985 DETEs from the 'shared' subset, 711 of them gained more activity with the addition of 4PBA than mock treatment alone (Appendix C.10 Figure C.15 F). These together explain the sources of 4PBA DETEs (Figure 7.9 F), in which the majority of the up and down-regulated DETEs were respectively contributed from expression candidates of '4PBA-unique' and 'mock-unique' subsets. As for DETEs from the expression candidates shared between the two treatments, most of them exhibited elevated activity in the presence of 4PBA.

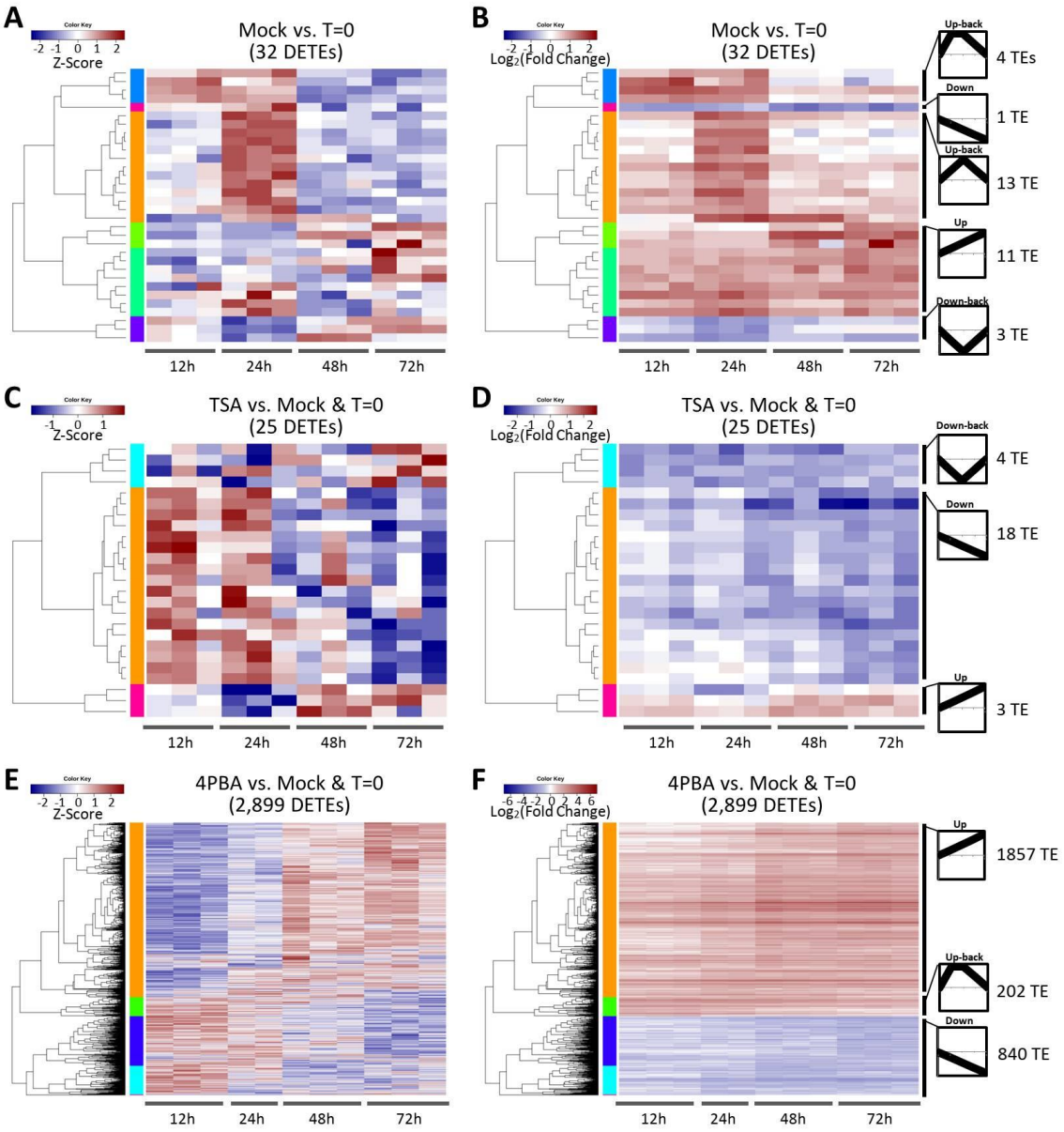



**Figure 7.9 Heatmaps and expression patterns of DETEs**

DETEs in mock, TSA and 4PBA treatments were illustrated by heatmaps using Z-score (A, C, E, respectively) and $log_2$(fold change) (B, D, F, respectively). The expression changes of clusters in heatmaps were interpreted into line graphs shown on the right side. The size of each cluster was indicated. As heatmaps of Z-score enhance the direction of expression changes (e.g. low to high or high to low across time) over the four time-points, $log_2$(fold change) emphasizes the changes comparing to T=0 or mock (e.g. activation or suppression in the specified treatment).

## 7.4.5 Investigation of autonomous TE transcripts

The transcriptional activity of the grapevine's autonomous LTR-TEs has been observed in embryogenic callus subjected to mock and biotic stress treatments, in which potential origins of autonomous LTR-TE transcripts were identified using Illumina sequencing data of chapter 3. Among the reported autonomous loci in chapter 3, Copia-3 and Copia-23 were the two LTR-TE families showing the most promising activity in all experimental condition. However, using ONT cDNA

sequencing, no competent TE read was detected, except for one validated read for the LTR-TE Gypsy-V1 and eight reads for DNA transposon hAT-7 in the mock treatment (see chapter 6).

Note that the 4PBA expression candidate pool contained a wider range of TE families than the mock expression candidate pool (Figure 7.2 D) and that a considerable number of trackable TE loci were differentially up-regulated in 4PBA treatment (Figure 7.9 E, F). Therefore, it was expected that a wider variety of autonomous expression candidates would be found in the 4PBA treatment than in mock treatment. Focusing on LTR-TEs, the same analysis approach, as described in chapter 3, identified 122 LTR-TE loci as potential origins of autonomous transcripts in 4PBA treated callus (Figure 7.10 B). The centre of the Venn diagram shows the 56 loci commonly found in the four experimental conditions (Figure 7.10 B). On the edge of the Venn diagram, nine autonomous LTR-TEs were unique to mock, whereas 34 loci were specifically found as the potential origins of autonomous TE transcripts in 4PBA treatments. In concordance with previous findings, the majority of these loci located in the introns of expressed genes (Figure 7.10 C). Although Copia-3 and Copia-23 are still the two TE family contributed most to this collection, there are other 26 LTR-TE families shown as potential origins of autonomous transcripts, 16 of which were only found in the 4PBA autonomous subset (Figure 7.10 D).

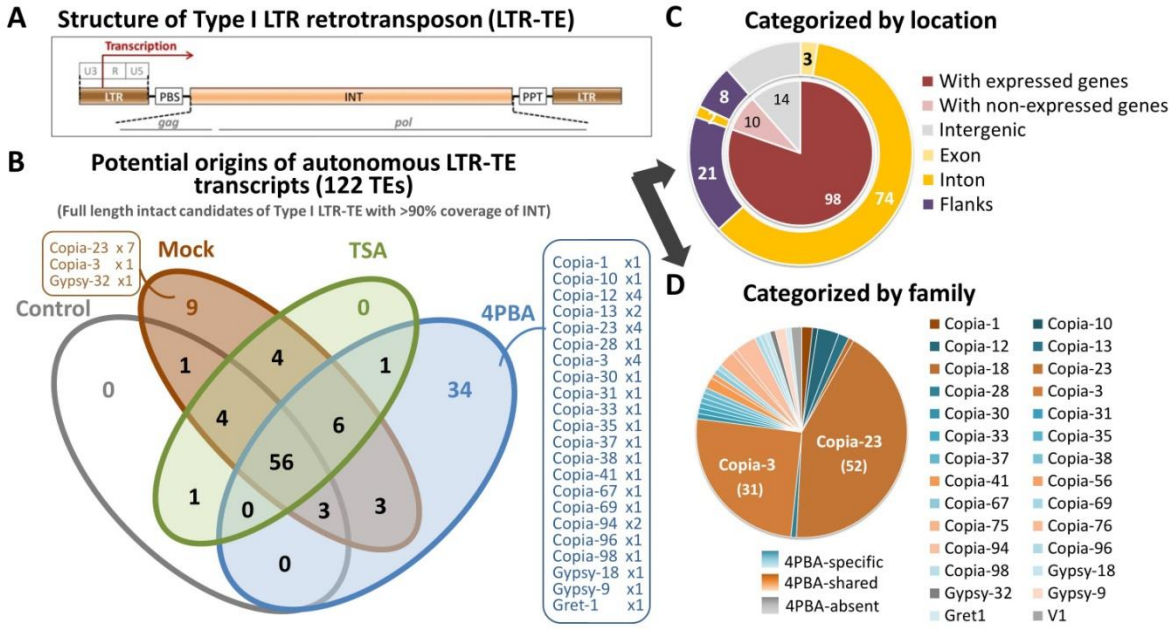



**Figure 7.10 Identification of potential origins of autonomous LTR-TE transcripts**

**(A)** Illustration of the canonical LTR-TE structure. Intact full-length LTR-TE loci with sequencing reads covering >90% of the internal domain (INT) were considered as potential origins of autonomous transcripts (see chapter 4 for more details). **(B)** Comparison of the autonomous LTR-TEs identified as the potential origins. There are 56 TE loci commonly found in all four conditions, while other 9 and 34 loci were unique to mock and 4PBA treatments, respectively. TE families contributed to the mock- and 4PBA-unique loci were listed. **(C)** Categorization of these loci by location. The number of TEs in each category was indicated. **(D)** Categorization of these loci by family. These TE families were further bin into three groups: uniquely found in 4PBA (4PBA-specific), shared with other conditions (4PBA-shared), and absent in the 4PBA autonomous subset (4PBA-absent). The number in the brackets denotes the number of TE loci of the corresponding TE family.

With the advantage of intact sequence information recorded in each sequenced read, ONT cDNA sequencing was conducted to validate the autonomous transcriptional activity of TEs. Considering the trend of transcriptional activation seen in the 4PBA samples, the full-length cDNA from 72h of the 4PBA treatment was sequenced, while samples of 72h of mock and T=0 were sequenced for comparison. Eighteen to 28 million reads were produced across replicates, corresponding to 12 to 19 billion sequenced bases (Table 7.3). After adapter removal and quality selection, 10 to 15 million reads, with N50 ranging from 703 bp to 737 bp, were aligned to the reference genome. Among these reads, 191 to 293,000 of them overlapped with annotated TEs. With such impressive sequencing depth, however, the overall N50 dropped from about 840 bp in the raw reads to the vicinity of 700 bp in the pool of mapped reads. Although the N50 indicates that the mapped pool seemed to be populated mostly by reads shorter than 1 kb, the longest mapped reads were found to be between 10 to 20 kb. Besides, with the degree of Pearson's correlation coefficient (ρ) at about 0.8, the gene expression quantification based on ONT data was highly correlated with the expression level obtained from the Illumina dataset (Appendix C.11 Figure C.16 A-C). For TE expression quantification at the family level, a medium to high level of correlation (ρ = 0.64 – 0.67) was obtained from the comparison between the ONT and the Illumina datasets (Appendix C.11 Figure C.16 D-F).

Following the workflow established in chapter 6, the transcriptional coverage of all annotated TEs was examined. Several TE loci had gained the full read coverage across the domains crucial for transposition (Appendix C.12). However, when further examined the alignment coverage on the genome browser, only very limited TE loci show the potential of competent transcription for autonomous mobilization. This includes a Gypsy-V1 locus having a unique-mapping ONT reads covering >90% of its INT domain in the 4PBA library (Figure 7.11) and six hAT-7 loci sharing one to two ONT reads (i.e. multi-mapping reads) that covering >90% of the ORF region (Figure 7.12).

**Table 7.3 Mapping statistics of oxford nanopore (ONT) cDNA sequencing (SQK-109)**

| | Sequenced | | | Adapter removal | | | Total mapped | | | TE-mapped | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Ctrl** | # Reads | 28,703,759 | 100.00% | # Reads | 20,874,513 | 77.59% | # Reads | 15,728,675 | 62.26% | # Reads | 293,976 | 1.08% |
| (00h) | Total bases | 19,361,506,482 | 100.00% | Total bases | 10,019,579,858 | 59.78% | Total bases | 9,011,383,396 | 55.13% | Total bases | 40,628,688 | 0.27% |
| | N50 | 831 | | N50 | 682 | | N50 | 703 | | N50 | 692 | |
| **Mock** | # Reads | 18,266,640 | 100.00% | # Reads | 13,715,437 | 75.08% | # Reads | 10,950,271 | 59.95% | # Reads | 191,037 | 1.05% |
| (72h) | Total bases | 12,474,922,541 | 100.00% | Total bases | 7,064,041,187 | 56.63 | Total bases | 6,429,997,000 | 51.54% | Total bases | 25,165,746 | 0.20% |
| | N50 | 843 | | N50 | 719 | | N50 | 737 | | N50 | 726 | |
| **4PBA** | # Reads | 21,622,640 | 100.00% | # Reads | 16,422,680 | 76.99% | # Reads | 13,064,612 | 64.00% | # Reads | 231,450 | 1.13% |
| (72h) | Total bases | 14,751,018,104 | 100.00% | Total bases | 8,419,898,861 | 60.33% | Total bases | 7,657,084,299 | 56.44% | Total bases | 29,538,982 | 0.28% |
| | N50 | 842 | | N50 | 717 | | N50 | 735 | | N50 | 723 | |

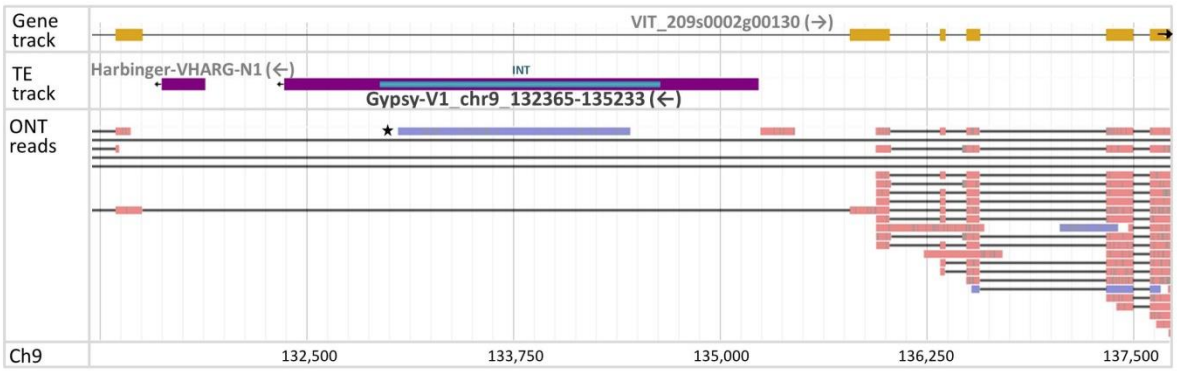



**Figure 7.11 Genome browser image of the autonomous Gypsy-V1 covered by ONT read.**

This TE locus, Gypsy-V1_chr9_132365-135233, is the only LTR-TE nearly fully covered by a single ONT read. It locates in an intron of VIT_209s0002g00130. The pink and blue strips denote forward and reverse reads, respectively. The black star marks the ONT read overlapping with >90% INT (teal blue strip) in the sense orientation. The orientations of annotated TEs and genes were as indicated.

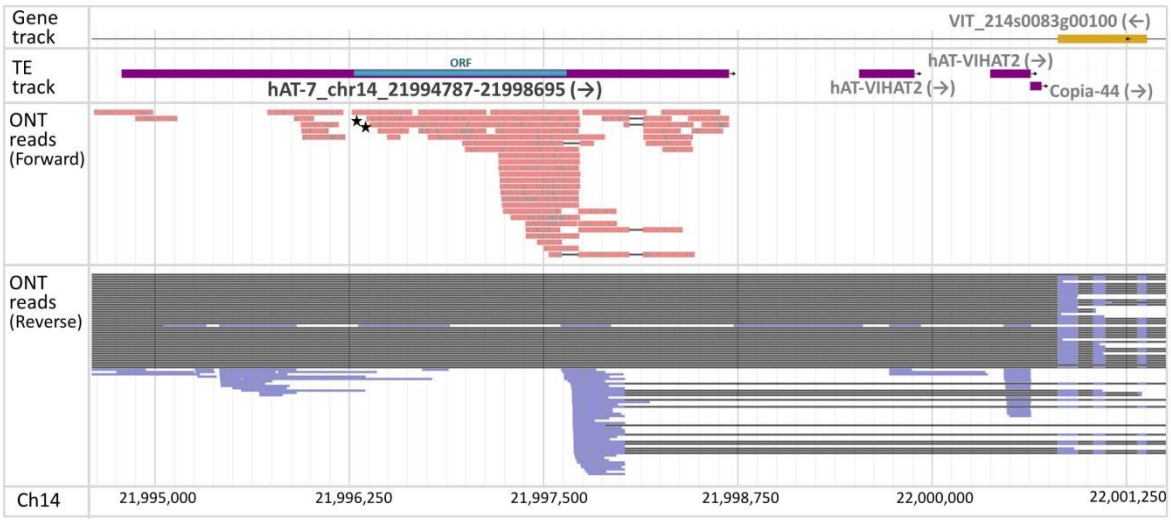



**Figure 7.12 Genome browser image of the representative autonomous hAT-7.**

This TE locus, hAT-7_chr14_21994787-21998695, is one of the three hAT-7 loci having ORF fully covered by individual ONT reads. The ORF (teal blue strip) identified using ORFfinder was overlaid manually and proportionally. It locates within an intron of VIT_214s0083g00100. The pink and blue strips denote forward and reverse reads, respectively. The single black star marks the ONT read overlapping with >90% ORF in the sense orientation. The orientations of annotated TEs and genes were as indicated.

### 7.4.6 Survey of stress-related CREs

The discordance between the transcriptional activation of TEs observed from the Illumina sequencing libraries and the lack of competent ONT cDNA reads for these TEs leads to the question that, with the presence of 4PBA, whether these TEs need an additional push from stimuli other than wounding. To investigate this possibility, a survey of stress-related cis-regulatory elements (CREs) was conducted. This analysis includes the 22 canonical LTR sequences (i.e. 22 LTR-TE families) representing the 34 potential autonomous loci unique to 4PBA treatments (Figure 7.10 B), as well as the canonical LTR sequence of Gypsy-V1 and the hTA-7 sequences prior to the ORF region. Wide varieties of stress-responsive CREs were found in these TE families (Figure 7.13 – 7.14). Although wound-responsive element, which is suggested to be the sort of CRE responding to the mock treatment that resembles wound stress (also see section 2.5.1), was commonly detected across these TE families, it was not the most prevalent one. The most predominant stress-responsive CRE varies depending on the surveyed TE families, but some of these predominant CREs were more common in multiple TE families than other CREs; these are CREs associated with pathogen, sugar, and heat responsiveness (Figure 7.13, Figure 7.14). Therefore, it is plausible that the supplement of 4PBA has tuned the permissive transcriptional area as covering a wider range of active TE families, yet it's not enough to boost the transcription for autonomous mobilization unless with the combinational application of the most promising biotic or abiotic stimuli.

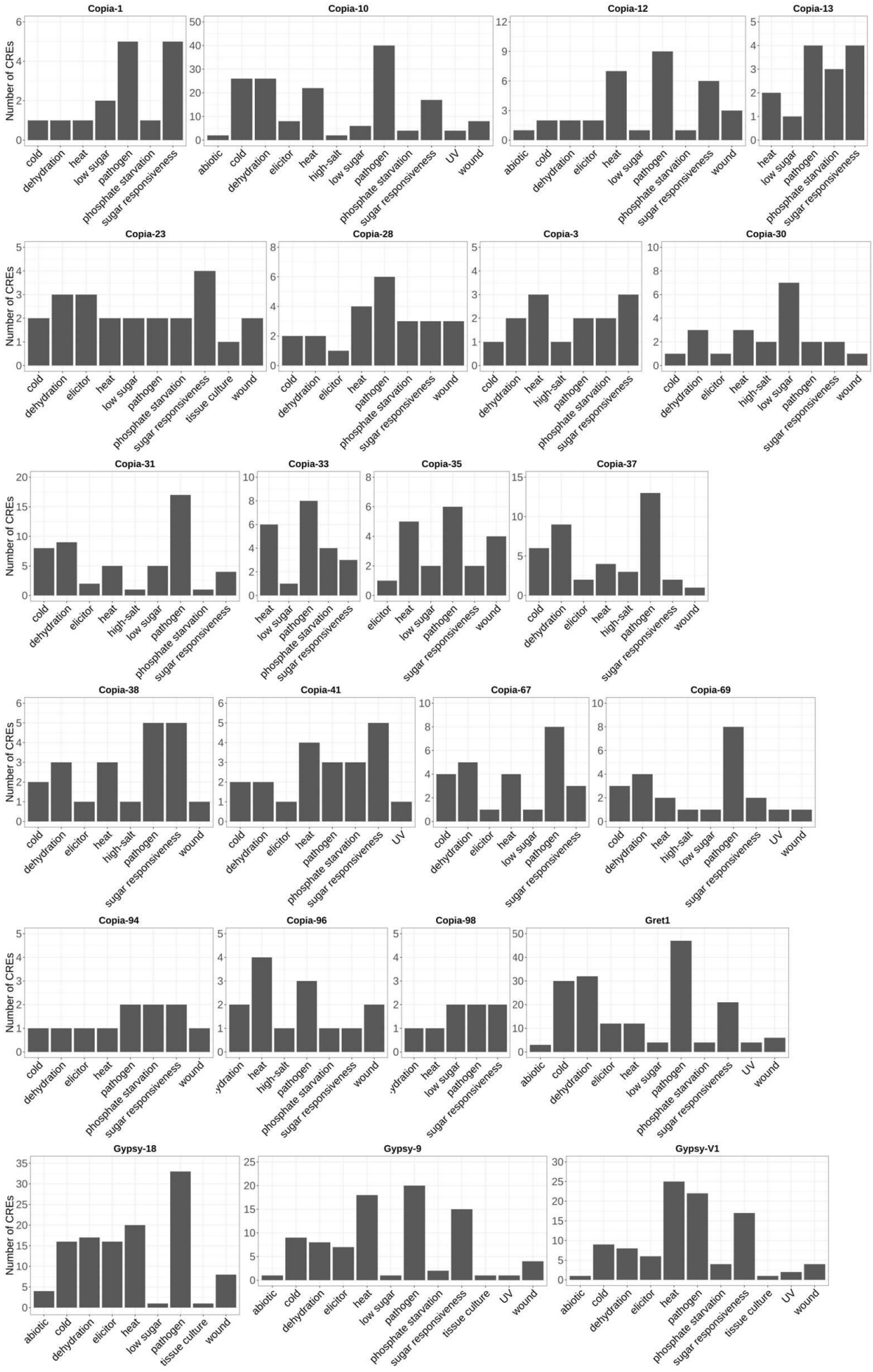

Copia-1
Copia-10
Copia-12
Copia-13
Copia-23
Copia-28
Copia-3
Copia-30
Copia-31
Copia-33
Copia-35
Copia-37
Copia-38
Copia-41
Copia-67
Copia-69
Copia-94
Copia-96
Copia-98
Gret1
Gypsy-18
Gypsy-9
Gypsy-V1
Number of CREs

**Figure 7.13 Survey of stress-related CREs of the LTRs from the selected LTR-TE families**

Twenty-three LTR-TE families were included in this analysis. The annotated CREs were grouped by their stress-responsiveness, as shown on the x-axis, while the counts of these CREs were projected on the y-axis.

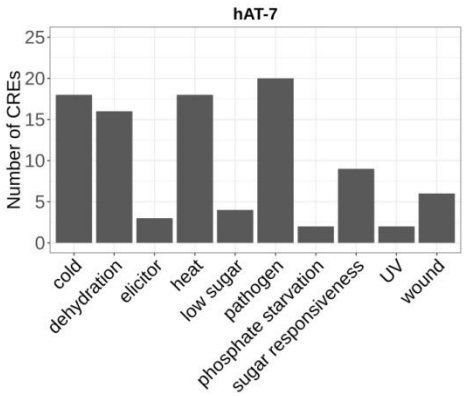



**Figure 7.14 Survey of stress-related CREs of canonical hAT-7**

The canonical hAT-7 sequences prior to the ORF was extracted and analysed. The annotated CREs were grouped by their stress-responsiveness, as shown on the x-axis, while the counts of these CREs were projected on the y-axis.

### 7.4.7 Expression pattern of grapevine genes potentially involving in the epigenetic machinery

In addition to the application of a less effective stressor like wounding, the multi-layered epigenetic system, which could catch transposition risks from multiple aspects, may also account for the underrepresentation of autonomous TE transcripts. To test this, grapevine genes similar to the Arabidopsis genes involving in epigenetic silencing were extracted from the data and examined for expression changes. Among the genes listed in Appendix C.9, 16 of them were differentially expressed in the statistical test of 4PBA versus mock datasets (Figure 7.15); while half were down-regulated, the other half were up-regulated. The roles of these 16 genes cover various silencing tasks and pathways, yet no pathway was entirely suppressed in the presence of the HDACi. For example, as some of the Argonaute proteins involving in post-transcriptional gene silencing (PTGS), e.g. AGO1 and AGO10 (Borges and Martienssen, 2015), were down-regulated, the AGO2 participating in both PTGS and RNA-dependent DNA methylation (RdDM), as well as AGO4 responsible for RdDM (Borges and Martienssen, 2015; Cuerda-Gil and Slotkin, 2016), were up-regulated. While the chromomethylase CMT2, which often targets the CHH sites governed by the chromatin remodeler DDM1 (Deniz et al., 2019), was lightly suppressed in the 4PBA sample, the other DNA methyltransferase DRM2, which involves in CHH methylation in the RNA Pol IV- or Pol V-dependent RdDM (Cuerda-Gil and Slotkin, 2016), was still expressed at a relatively stable state (i.e. not

differentially expressed). Noticeably, the expression of the Dicer gene DCL2 was increased at 48h and 72h post the mock treatment, yet had a four-fold down-regulation with the continuous presence of 4PBA. DCL2 is functionally redundant with DCL4 as both of them participate in the first layer of defence against transcriptional TE activation (section 1.3.2; Cuerda-Gil and Slotkin, 2016; Marí-Ordóñez et al., 2013). Although DCL2 was down-regulated in 4PBA treatment, there was no difference in the expression level of DCL4 between mock and 4PBA treatments. Overall, although these down-regulated genes possess various specificities in epigenetic silencing pathways for different targets, the up-regulated and the unaffected genes that were functionally redundant or involving in the same pathway as the down-regulated genes might still keep the epigenetic silencing networks competent in suppression of TE activity.

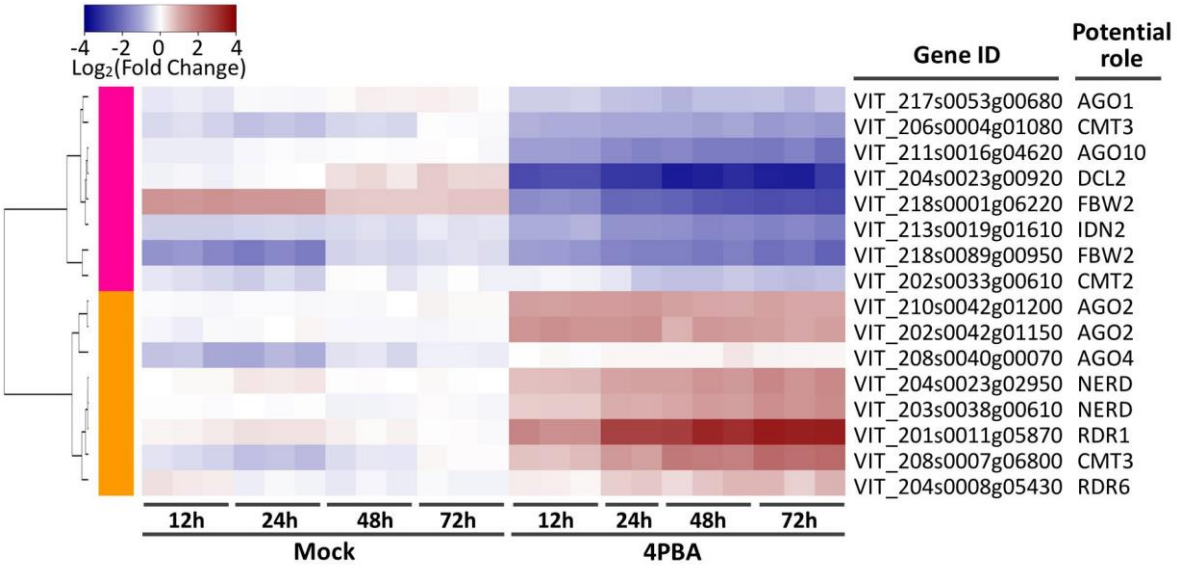



**Figure 7.15 Heatmap of differentially expressed genes in 4PBA treatment potentially having epigenetic roles in grapevines**

DEGs in the 4PBA treatment potentially with epigenetic roles were illustrated by heatmap using $log_2$(fold change). The $log_2$(fold change) was calculated by normalizing gene expression level in mock and 4PBA samples to the control (T=0). Therefore the fold change displayed in white colour represents the expression level as same as T=0. The gene ID and the potential roles were shown on the right side for each row in the heatmap.

## 7.4.8 Gene ontology analysis for the 4PBA-treated samples

Among the 5678 DEGs in 4PBA treatment, 3482 genes were up-regulated, and the rest were down-regulated compared to mock over time (Appendix C13 Figure C.17). Although 61% of the DEGs in 4PBA treatment were up-regulated compared with mock treatment, the enriched gene ontology (GO) network represented by these up-regulated DEGs only comprised ten nodes related to biological process (Figure 7.16 A) as well as four nodes for cellular component and molecular function networks (Figure 7.16 B, C). In the up-regulated biological process network (Figure 7.16 A),

nodes corresponding to the response to chitin, hydrogen peroxide, high light intensity, and heat acclimation were linked together, while the node of toxin catabolism was associated with a node representing nuclear-transcribed mRNA catabolism, and the node of peroxisome organization was linked with a node of membrane fusion. There were two nodes independent from others; these were node representing the response to bacterium and node related to biphenyl metabolism. The enriched GO nodes in the cellular component network were associated with the endoplasmic reticulum (ER) lumen and nucleus (Figure 7.16 B), while the nodes in the molecular function network were corresponding to aryl-alcohol dehydrogenase activity and abscisic acid glucosyltransferase activity (Figure 7.16 C). These up-regulated GO nodes suggest increased oxidative stress (Demidchik, 2015; Takahashi and Murata, 2008) in the grapevine embryogenic callus subjected to 4PBA treatment. Interestingly, the cellular component network regarding ER lumen was enhanced (Figure 7.16 B). This supports the findings in mammalian neuronal cells suffering from ER stress, in which ER homeostasis was rescued by 4PBA treatment (Kusaczuk et al., 2015).

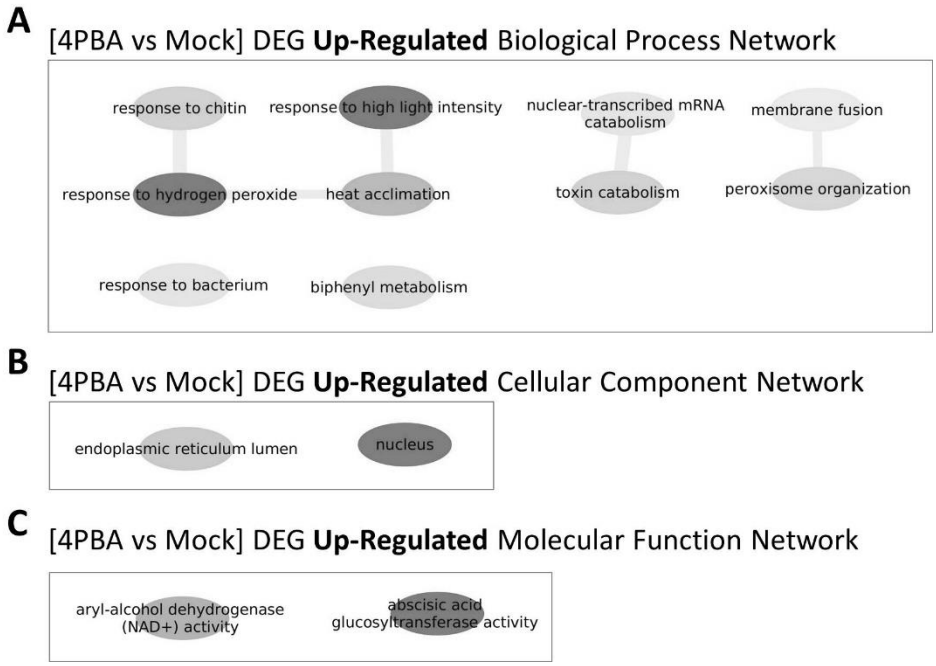



**Figure 7.16 Enriched GO networks of up-regulated DEGs in 4PBA treatment**

Significant GO terms ($p < 0.05$) in **(A)** biological process, **(B)** cellular component and **(C)** molecular function networks for up-regulated DEGs in 4PBA treatments. Links denote closely related GO term clusters, among which the darker, the more significantly enriched (lower REVIGO p-value).

Given the 40% down-regulated genes among the DEGs, the significantly suppressed networks were much more complicated than the activated series of DEGs (Figure 7.17, Appendix C14). The down-regulated biological networks mainly comprised responses to biotic and abiotic stresses (Figure 7.17 A), regulation of cell differentiation and tissue development (Figure 7.17 B), as well as cell wall

organization and transport of water, glucose and toxin such as organophosphate ester (Figure 7.17 C). These three groups were connected through signalling pathways involving protein phosphorylation, ubiquitination, methylation and possibly acetylation (Figure 7.17 D). The interconnecting relationships between cell wall organization, plant tissue development and stress response have been reported previously (Flematti et al., 2015; Luan, 2002; Tenhaken, 2015). A smaller group of down-regulated networks are related to salicylic acid metabolism (Figure 7.17 E), which is considered to be closely associated with biotic and abiotic stress response (Lefevere et al., 2020). The rest of the clusters are isolated from others. However, they can be linked with the aforementioned groups from the perspectives of their biological meanings in flavonoid biosynthesis and cell cycle regulation (Chalker-Scott, 1999; Qi and Zhang, 2020). The down-regulated networks regarding cellular component and molecular function are consistent with those of biological process (Appendix C14 Figure C.18).

The up-regulated GO networks are generally related to the down-regulated networks., For instance, up-regulated response to high light intensity (usually associated with increased oxidative stress and photoinhibition) may cause reduction of plant growth (Demidchik, 2015; Goh et al., 2012), up-regulated toxin catabolism may be in response to the halted organophosphate ester transport (Liu et al., 2019), heightened activity of abscisic acid glycosyltransferase can reduce drought tolerance (Liu et al., 2015), and aryl-alcohol dehydrogenase (NAD+) activity may be negatively related to MAPK signalling pathway.

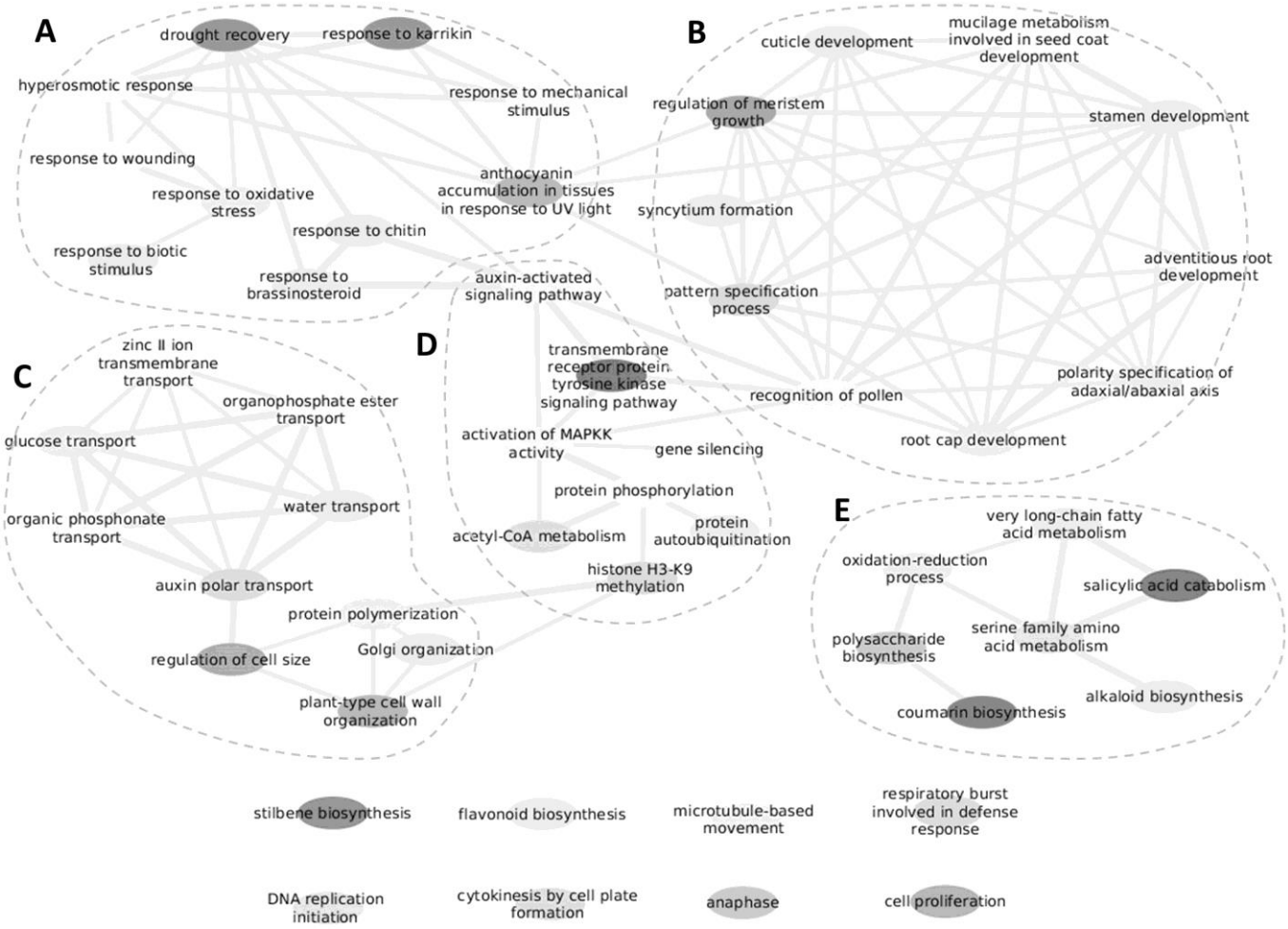



**Figure 7.17 Enriched biological process networks of down-regulated DEGs in 4PBA treatment**

For down-regulated DEGs in 4PBA treatments, significant GO terms ($p < 0.05$) in biological process networks can be sorted into five groups: **(A)** responses to biotic and abiotic stresses, **(B)** regulation of cell differentiation and tissue development, **(C)** regulation of cellular components and transportation, **(D)** signalling pathway involving protein modification, and **(E)** salicylic acid metabolism. Some clusters isolated from others were displayed at the bottom. Links denote closely related GO term clusters, among which the darker, the more significantly enriched (lower REVIGO p-value).

## 7.5 Discussion

### 7.5.1 Differences between TSA and 4PBA in terms of TE activation

Based on the current data, the application of TSA to grapevine embryogenic callus didn't show significant differences in terms of TE loci transcriptional activation when compared to mock treatment. It seems that the TSA treatment didn't stimulate TE activity. In contrast, the 4PBA treatment resulted in a significant shift in transcription pattern with a total of 5,894 differentially transcribed TE loci compared with mock treatment (Figure 7.2 C); 2,059 and 840 of these TE loci were found to be significantly up-regulated and down-regulated, respectively (Figure 7.9 F). It's apparent that 4PBA has stronger effects than TSA in terms of TE activity in the given experimental design. In addition, grapevine embryogenic callus subjected to TSA exhibited browning in 12 hours of treatment, whereas mock and 4PBA treatments did not show such effect. This suggests that the differences between TSA and 4PBA in terms of TE activation might be associated with the cytoprotective properties of 4PBA (section 7.2.3; Kusaczuk et al., 2015) and cytotoxicity of TSA (Alao et al., 2006; Blagosklonny et al., 2002, 2005). Blagosklonny et al. (2002) investigated the cytotoxicity of three classes of HDACi: sodium butyrate (structurally similar to 4PBA but deprived of phenyl group), TSA, and depsipeptide (specifically inhibits human HDAC1 and HDAC2; Bolden et al., 2006; Kusaczuk et al., 2015). Their results show that the cytotoxicity of sodium butyrate is the least among these three HDACi in human cancer cell lines. In addition, as previously mentioned, the butyrate derivatives containing phenyl group, such as 4PBA, were found to exhibit cytoprotective properties. These together might explain the differences in TE transcriptional activity between grapevine embryogenic callus treated by TSA and 4PBA.

### 7.5.2 Regionally specific responsiveness to 4PBA treatment

The observations in section 7.4.2 to section 7.4.3 suggest a shift of transcriptionally permissive areas rather than genome-wide broad expansion of the active areas in the 4PBA-treated samples. As mentioned in the introduction, short-chain fatty acid-type of HDACi, such as butyrate and its chemical derivative 4PBA, can specifically inhibit the acetylation activity of human class I and some class II HDACs (Bolden et al., 2006; Kusaczuk et al., 2015). Although its specificity to various types of plant HDAC is currently unclear, it is possible that 4PBA only suppressed the activity of some grapevine HDACs, and thus resulted in regional activation of TEs. In fact, this speculation was strengthened with the analysis of location distribution of expression candidates, in which elevated proportion of intergenic and flanking-region located TE loci were seen in the 4PBA-treated samples compared with mock treatment. Besides, these changes were mostly contributed from the

expression candidates present in the candidate pool of 4PBA treatment but absent in the candidate pool of mock treatment (Figure 7.5, Figure 7.6). In addition, it has been reported that different HDACs act on various genes with different specificity and that HDACs only interact with specific chromatin regions (Kurdistani and Grunstein, 2003; Kurdistani et al., 2002; Wang et al., 2009). It is possible that the intergenic TE loci uniquely identified as expression candidates in 4PBA treatment might be closer to expressed genes than was the intergenic expression candidates present in mock treatment but absent in the candidate pool of 4PBA treatment. Indeed, this assumption was supported by the result shown in Figure 7.7. By resolving the genome-wide binding map of the yeast class I HDAC Rpd3, Kurdistani et al. (2002) found that Rpd3 was preferentially associated with promoters of highly expressed genes and was absent from sub-telomeric regions. Similar binding preference towards gene promoters was also observed for the yeast class II HDAC Hda1, albeit differences existed between Rpd3 and Hda1 in terms of target gene specificities as well as targeted histones (Kurdistani and Grunstein, 2003). In human T cells, Wang et al. (Wang et al., 2009) found that multiple types of HDACs predominantly bind with chromatin regions enriched with expressed genes, as some of the HDACs preferentially target the promoter or enhancer regions and the others target the transcribed regions. Their results revealed that HDAC's binding hot spots highly overlapped with those of RNA Pol II and HAT, in addition to its positive correlation with histone acetylation and gene expression levels. Moreover, these human HDACs were absent in the chromatin regions lacking methylated H3K4, a hallmark indicative of transcriptional permissiveness (Wang et al., 2009). Genes with H3K4 methylation were considered inducible yet not necessarily active constantly. Coupled to increased binding of Pol II at the promoters, Wang et al. (2009) found that inhibition of HDAC activity with TSA and butyrate increased histone acetylation level at genes with H3K4me3, instead of genes without H3K4 methylation. These results show that the histone acetylation level of active genes is maintained by sophisticated coordination of HAT and HDAC, while the adding of the acetyl groups by the former is faster than the removal by the latter. Although inhibition of HDAC can further increase the acetylation level, it is proposed that this may increase chromatin instability and chances of aberrant transcription (Wang et al., 2009). On the other hand, inactive genes with H3K4 methylation interact with HAT and HDAC at a lower frequency than the active ones. When the HDAC activity is compromised, these genes can acquire the accessibility to Pol II when the HDAC activity is compromised, albeit their transcription remained undetected (Wang et al., 2009). This suggests that inactive genes primed by H3K4 methylation can be ready for transcriptional activation upon receiving stimulus signals. As for the silenced genes lack of H3K4 methylation, their interaction with HAT and HDAC was found to be undetectable, and therefore remained inactive when HDAC was inhibited (Wang et al., 2009).

The studies in yeast and mammals fit with the findings in the 4PBA-treated grapevine callus, where an increase in the proportion of expression candidates in the flanking regions of active genes was observed. Furthermore, the newly emerged intergenic expression candidates in 4PBA treatment were not randomly distributed in the intergenic region but showed a correlation with expressed genes in terms of their physical distance along the chromosome. It is sensible to speculate that these intergenic and flanking regions of expressed genes contain the corresponding promoters and enhancers, where the histone acetylation levels may have been elevated by the HDACi 4PBA. Although the active TEs in these regions might serve as part of the promoters or enhancers for the gene expression, their detectable transcriptional activity more likely suggests that they took advantage of the permissive transcriptional condition. The aforementioned HDAC specificity in favour of regions primed by H3K4 methylation may also explain why there was no genome-wide scale of TE activation that could be reflected on the total number of expression candidates comparing to mock.

### 7.5.3 Rarely detected evidence of competent transcription from autonomous TEs albeit the noticeable transcriptional up-regulation

From the Illumina sequencing data, the noticeable uniqueness of the expression candidates and the up-regulated activity at individual TE loci in 4PBA treatment suggested elevated chances of competent TE transcriptions that may lead to autonomous transposition. While the Illumina sequencing data supported this possibility by showing an increased number of autonomous TE loci potentially accounted for the competent transcription in the 4PBA samples, the ONT cDNA sequencing data revealed a very limited number of the competent reads. A similar contradiction has been observed in the case of biotic stress treatment (see chapter 4 and chapter 6), in which the promising activity of autonomous TE loci seen in the Illumina dataset was not reflected by the ONT cDNA sequencing data. These suggest that neither the biotic stress nor the inhibition of HDAC alone was sufficient to trigger autonomous TE mobilization.

There are several reasons that may explain the lack of intact cDNA reads from autonomous TEs, given that the transcriptional activity and the up-regulation observed through the time series was clearly detectable in the short-read sequencing data.

Firstly, in the Illumina data, the transcriptional activation associated with the autonomous TE loci may be misleading due to the fragmented multi-mapping of TE transcripts derived from degenerated TEs or the interference from TE transcription with aberrant transcriptional starts or premature stop sites. With the library preparation for Illumina sequencing, the transcript information was processed into small pieces of cDNA fragments that cannot help but undermine the resolution of this approach.

Secondly, provided a TE transcript is part of the transcription unit driven from the host gene, this gene-TE fusion transcript would be subject to the nonsense-mediated mRNA decay (NMD) even if it is polyadenylated (Valencia-Sanchez et al., 2006). These kinds of transcripts are usually unstable and present at a much lower frequency in the transcriptome than normal transcripts, thus less likely to be captured during the preparation of the ONT cDNA library.

Thirdly, the epigenetic silencing system remains intact in our system despite our attempts at disrupting the re-formation of the heterochromatic structure by inhibiting the activity of HDAC. Although some of the genes involving in RdDM and PTGS were down-regulated in the presence of 4PBA (Figure 7.15), others were up-regulated or remained unchanged. From these data, it seemed that the epigenetic silencing network was not impaired with compromised HDAC activity, and therefore the induced level of TE activity was likely captured and contained by the surveillance machinery.

Last but not least, the competent transcription from autonomous TEs might have actually taken place and escaped from PTGS, yet these transcripts were expressed at a low level such that it evades detection using the ONT cDNA method. It is unclear that how many copies, or at what proportions, autonomous TE transcripts are required to be present in the total transcriptome for the detection using the ONT cDNA method. For cases of LTR-TEs that proliferate through the formation of the virus-like-particles (VLPs), some hints might be indicated from the detection of human immunodeficiency virus type 1 (HIV-1) RNA packaging in the VLPs. Using human serum spiked with HIV virus particles and clinical plasma from HIV patients, Erice et al. (2000) found that 50 copies of the virus RNA per mL were the detection limit in their assay. With a more sensitive technology, it is reported that quantitative PCR (qPCR) is able to detect the cDNA derived from 10 copies of RNA (Ferrer et al., 2016). If this scenario was the case in our situation, the expression level inferred from the short-read RNAseq might represent that the TE activity competent for mobilization is similar to the minimum detection level of ONT technology even with the biotic stimuli or pharmacological inhibition of HDAC. This scenario in respect of the low level of TE expression is concordant with the observations in the natural population of diploid and autotetraploid *Arabidopsis arenosa*, which reveals low-frequency accumulation of TE insertions occur in nature (Baduel et al., 2019). These TE insertions were found to be subjected to relaxed purifying selection while their deleterious effects can be masked by the high heterozygosity of the natural population (Baduel et al., 2019). This study suggests that the transcriptional activity of the active TEs is likely low but allows the accumulation of new TE insertions at low frequency.

### 7.5.4 Multiple effects of 4PBA

By suppressing HDAC activity, the application of 4PBA can have broad biological effects depending on the functionality and target specificity of inhibited HDAC. 4PBA is considered as an inhibitor to the class I HDAC, which includes those homologous to the yeast RPD3 protein. In *Arabidopsis*, the activity of class I HDAC AtHDA19 has been reported to be responsible for the development of shoot apical and root apical meristem (Liu et al., 2019), while another Class I HDAC AtHDA6 was found involving in flower development (Yu et al., 2011). Both AtHDA6 and AtHDA19 were reported to be crucial for salt and drought tolerance (Chen and Wu, 2010), light responsiveness (Jang et al., 2011; Tessadori et al., 2009), as well as pathogen responsiveness (Devoto et al., 2002; Zhou et al., 2005). In addition to removing the acetyl group from histone, HDACs have non-histone substrates involving in a wide range of signalling pathways and biological functions. In other words, the function and activity of these substrates can be regulated by the deacetylation activity of HDACs. So far, the identified non-histone substrates in humans include the transcription factors p53 and E2F3, cytoplasmic heat shock protein Hsp90, DNA repair subunit Ku70, as well as structural proteins α-tubulin and β-catenin (Ma et al., 2013). These studies explain some of the biological process networks that we found to be significantly down-regulated in the 4PBA treatments, such as networks related to drought recovery, regulation of meristem growth, anthocyanin accumulation in tissue in response to UV light, response to biotic stimulus, cell cycle regulation, protein phosphorylation, as well as the plasma membrane and cell wall organization (Figure 7.17).

HDACs are also found to participate in the regulation of epigenetic silencing. Mutations in the *AtHDA6* gene were associated with a reduced level of DNA methylation at CG and CHG (Earley et al., 2010) and increased transcriptional activity of a subset of TEs and rRNA genes (Earley et al., 2006; Yu et al., 2017). In addition, it has been reported that AtHDAC6 protein interacts with DNA methyltransferase MET1 (Liu et al., 2012; To et al., 2011b) and histone methyltransferases SUVH4, SUVH5 and SUVH6 (Yu et al., 2017). These existing findings are consistent with the down-regulated GO networks associated with histone H3K9 methylation and gene silencing in 4PBA treated sample (Figure 7.17 D). Indeed, our research is the first time to define the transcriptional effect of pharmacological inhibition of HDACs on the expression of genes participating in the epigenetic silencing pathway. In our data, distinct differential expression patterns were observed between grapevine genes that potentially participated in PTGS and RdDM (Figure 7.15). This finding is further discussed as follows.

Evidence from literature shows that AGO1, AGO2, DCL2, DCL4, and RDR6 are key factors acting at the front line of TE activation to reduce increased TE transcripts (Borges and Martienssen, 2015; Cuerda-Gil and Slotkin, 2016; Marí-Ordóñez et al., 2013). AGO1 has been considered as a key component

required for the PTGS pathway, in which it carries 21-22 nt siRNAs and miRNAs to mRNAs that contain sequences complementary to the siRNAs or miRNAs and leads to cleavage of the mRNAs (Creasey et al., 2014; Cuerda-Gil and Slotkin, 2016). The function and activity of AGO1 were found to be partially redundant with that of AGO2 for that a small subset of miRNAs (e.g. miR165 and miR408 in *Arabidopsis*) can bind with both AGO1 and AGO2 (Bologna and Voinnet, 2014; Borges and Martienssen, 2015). *Arabidopsis* DCL2 and DCL4 process long double-stranded RNAs into 22nt and 21nt siRNA, respectively (Bologna and Voinnet, 2014; Borges and Martienssen, 2015). Despite differences in size, both siRNAs generated by DCL2 and DCL4 can achieve a similar effect of PTGS (Borges and Martienssen, 2015; Cuerda-Gil and Slotkin, 2016). RDR6 uses AGO1-preprocessed TE transcripts as templates to generate double-stranded RNAs (dsRNAs), which are then diced into 21-22nt secondary siRNA by DCL2 and DCL4. In our system, the PTGS pathway affected by down-regulation of AGO1 and DCL2 in 4PBA treatment might not be fully compensated with unaffected DCL4 expression and up-regulated AGO2 and RDR6, because AGO2 only partially redundant with AGO1 as well as RDR6 acts downstream of the primary 21-22nt siRNA synthesis by DCL2/DCL4 and mRNA cleavage by AGO1 (see Figure 1.3 for the illustrated PTGS pathway). However, the up-regulation of AGO2 and RDR6 might in turn enhance a non-canonical RdDM pathway described as follows.

From studies of the plant model system *Arabidopsis*, Pol IV-NERD-mediated RdDM has been reported as a non-canonical RdDM pathway, where the long dsRNAs generated by Pol IV, RDR6 and RDR1 can be processed by Dicer proteins into siRNAs, which are then loaded onto AGO2 and carried to Pol V-derived transcripts with help from NERD protein (Need for RDR2-independent DNA methylation; Cuerda-Gil and Slotkin, 2016). The interaction of the siRNA-bounded AGO2, Pol V, and NERD induce the recruitment of DNA methyltransferase to deposit methylcytosine (Cuerda-Gil and Slotkin, 2016). Intriguingly, we found that not only AGO2 and RDR6 but also RDR1 and NERD were significantly up-regulated in 4PBA treated sample. Therefore, in our system, it is possible that this non-canonical RdDM pathway was enhanced in 4PBA treatment. We also found another Argonaut protein, AGO4, was up-regulated in 4PBA treatment. AGO4 is believed to be a necessity of the canonical RdDM pathway, where it carries 24nt siRNA to Pol V-produced transcripts, facilitating the recruitment of DNA methyltransferase to complete the RdDM pathway (see Figure 1.2 for illustrated pathway; Cuerda-Gil and Slotkin, 2016).

Overall, although the PTGS pathway seems to be negatively affected, it appears that the canonical and non-canonical RdDM pathways might be enhanced in grapevine embryogenic callus subjected to 4PBA treatment. In wild-type rice, transcriptional activation of TE loci in response to phosphate starvation was found to be followed by re-enhancement of DNA methylation on these sites (Secco et

al., 2015). In *Arabidopsis*, it has been reported that the host cells can suppress TE activity by maintaining or enforcing RdDM pathways on the transcriptionally active TE sites (Panda et al., 2016; Sigman and Slotkin, 2016). Therefore, in our system, the enhancement of canonical or non-canonical RdDM pathways (implied by up-regulated AGO2, AGO4, RDR1, RDR6 and NERD) might be a 'reflex' of host cells to compensate the impact on the PTGS pathway or to suppress the fluctuation of TE transcription in the presence of 4PBA.

### 7.5.5 Factors for massive autonomous TE activation in wild-type embryogenic callus

The discussion above (section 7.5.3) shows that it requires multiple factors to achieve the effective stimulation of autonomous TE activity in wild-type backgrounds. The right pick of environmental stimuli plays an important role in TE activation. Our analysis of the CREs housed within the transcriptionally active autonomous TEs shows that most of the investigated TE families are enriched with pathogen, sugar or heat responsive elements (Figure 7.13 – 7.14). Although some of these TEs were activated in the wound-like treatment, the wound-responsive CRE did not appear to be the most prevalent. On consideration of the identified CREs, stressors such as heat shock might stimulate TE activity more efficiently than the stresses link with the less prevalent CREs.

However, it seems that the suitable environmental cue alone is not sufficient to boost the TE transcription that is competent for autonomous mobilization. In the study in grapevines, the elevation of TE transcriptional activity in response to biotic stresses was detected by qPCR (Lizamore, 2013) and short-read RNAseq (chapter 4). However, full-length transcript derived from autonomous TE loci was hardly detected in the ONT cDNA data. Increased level of 21- to 22-nt siRNA was reportedly accompanied by TE activation (Borges et al., 2018; Marí-Ordóñez et al., 2013), indicating that PTGS can diminish the full-length TE transcripts if it is not swamped by large scale TE activation. Therefore, combined use of environmental cue and pharmacological inhibition of key enzymes modulating chromatin structure, PTGS and RdDM might give rise to transposition burst in the wild-type background. Recently, inhibitors for DNA methyltransferase and RNA Pol II have been applied on wild-type *Arabidopsis* and Rice underwent heat stress, leading to a more extensive burst of *ONSEN*'s retrotransposition than applying heat shock alone (Thieme et al., 2017).

Taken together, the findings in this chapter give rise to the hypothesis that the application of HDACi, in this situation 4PBA, acts to alter the chromatin landscape around genes that are becoming transcriptionally active due to perception of stress through preventing histone deacetylation of these areas. In this scenario, TEs close to or within the genes preferentially affected by 4PBA acquired the relaxed chromatin status and more frequent access to the transcription machinery, and thus resulted

in a wider range of potentially active TEs detected by Illumina RNAseq. However, without proper stress stimuli, any sporadic TE transcription would not be able to saturate PTGS machinery; instead, the fluctuation in TE transcription may, in turn, trigger stronger RdDM since the potentially active TE loci are not novel to the host cells. Future challenges in our system will be to interrogate whether a TE burst can be induced by drug inhibition of HDAC, DNA methyltransferase or RNA Pol II activity coupling with suitable stressors (e.g. heat) reflecting the presence of matching CREs that are found to be enriched in potentially autonomous TE loci associated with regions of chromatin affected by HDACi.

## 7.6 Conclusions

From the perspective of TE activation, the HDACi 4PBA is more effective than TSA in grapevine embryogenic callus. The 4PBA treatment did not increase the total number of potentially expressed TE loci; rather, 4PBA treatment resulted in the repression of 3,263 TE loci and activation of an alternate set of 2,653 TE loci, while another 3,985 TE loci present in the expression candidate pool of 4PBA treatment also existed in the candidate pool of mock treatment. This was coupled with a wider range of TE families found contributing TE loci to the 4PBA expression candidate pool. In comparison to the control samples (T=0) and mock-treatment, the 4PBA treatment resulted in an increased proportion of TE expression candidates located in intergenic as well as flanking regions of genes. This shift in distribution can mainly be attributed to the expression candidates uniquely identified in the 4PBA treatment. These implicate that within the confines of the time period of the experiments, 4PBA altered the landscape of TE activity, not by globally broadening transcriptionally permissive regions but by shifting the spectrum of the transcriptionally permissive area. Besides, the transcriptional activity of genes seems to be one of the determinants of the breadth of the permissive area for TE transcription. Transcriptionally activated genes and TEs might reflect the target specificities of 4PBA and the HDACs inhibited by 4PBA. It has been known that 4PBA specifically inhibit class I HDAC in mammalian systems. Moreover, existed evidence shows that HDACs preferentially bind to chromatins that associate with highly active genes as well as inactive genes primed by H3K4 methylation, which is associated with the ability of genes to be inducible under conditions that inhibit HDAC activity (Kurdistani and Grunstein, 2003; Kurdistani et al., 2002; Wang et al., 2009). Sub-telomeric and heterochromatic regions that lack H3K4 methylation display no interaction with HDAC; thus, the transcriptional permissiveness in these areas was very limited, even with the presence of HDACi. From the gene ontology analysis of DEGs in the 4PBA treatment, various networks have been significantly affected. These together may explain the regional instead of global activation of TE activity in the presence of HDACi 4PBA.

While multiple potentially autonomous TE loci became transcriptionally active under these treatments, ONT sequencing of full-length transcripts revealed few loci that were producing transcripts competent for mobilization. This could be due to several scenarios, such as misguidance of short-sequencing reads derived from degenerated TE loci, TE-gene fusion transcripts vulnerable to nonsense-mediated degradation (eliminating mRNA transcripts containing premature stop codons), still functional PTGS and RdDM systems, or a situation where the level of full-length TE transcripts was below the detection limit of ONT platform. To overcome these in the wild-type background, pharmacological inhibition of the PTGS and RdDM machinery combined with a suitable selection of stresses according to the prevalence of stress-responsive CREs appeared in TEs might give rise to a

significant transposition burst. Furthermore, methods like VLP-based analysis (Lee et al., 2020) and ONT sequencing platform are likely to be more capable of revealing active and competent TE loci.

# Chapter 8

# Analysis of small RNA dynamics in grapevine embryogenic callus exposed to pharmacological inhibitors of histone deacetylase

## 8.1 Overview

Previously, HDACi incubation in our embryogenic callus system resulted in the transcriptional activation of a wide range of TE loci. However, no full-length TE transcripts derived from autonomous loci were detected by the ONT cDNA sequencing dataset. This led to the speculation that the epigenetic silencing system was still competent in diminishing TE activity, despite inhibition of HADCs. We had also observed a down-regulation of AGO1 and DCL2 that participate in post-transcriptional gene silencing (PTGS) and an up-regulation of factors (e.g. AGO2, AGO4, RDR1, RDR6, and NERD) involving in canonical and non-canonical RNA-dependent DNA methylation (RdDM) pathways in 4PBA treatment, suggesting that RdDM, instead of PTGS, was enhanced in 4PBA treatment. In this chapter, examination of small RNA populations from the same grapevine embryogenic callus showed that the accumulation level of miRNAs and siRNAs targeting or derived from TEs was generally stable across treatments over time. In fact, 98 TE families (including Copia-3 and Copia-23) were found continuously seeding and maintaining TE-derived 24 nt siRNAs at levels > 100 RPM, suggesting the strong epigenetic suppression from the RdDM pathway onto these TE families. We further identified that the most highly and stably expressed miRNA, vvi-miR159c, was predicted to target Copia-23 and gene VIT_211s0016g05010, which potentially encodes Glyoxalase I 7 that is crucial for stress tolerance in plants. Intriguingly, in 4PBA treatment, the expression level of VIT_211s0016g05010 was significantly increased, whereas Copia-23 expression was two-fold inhibited compared with mock treatment. Although both this gene locus and Copia-23 are potentially targeted by the highly expressed vvi-miR159c, epigenetic suppression on this gene locus might be relaxed due to the negatively affected PTGS pathway observed in chapter 7, whereas the epigenetic suppression on Copia-23 was likely enhanced predominantly due to strengthened RdDM circuit. With these findings, we hypothesized that, with a sporadic or non-threatening level of TE activation, the 24 nt-siRNA-mediated RdDM is the major force to re-enhance silencing on the long-existing TEs in the genome. By contrast, PTGS may, in turn, be de-escalated to allow up-regulation of genes important for stress tolerance (e.g. VIT_211s0016g05010 in our case).

## 8.2 Introduction

The multi-layered epigenetic system responds to TE activation by triggering sequential networks, such as small RNA biogenesis and targeting, DNA methylation, as well as histone modification. The various outputs from different components of these networks act in combination to fine-tune gene or TE activity when cells are confronted with specific developmental stages or external stimuli. For this purpose, many different types of histone modifications exist, and each type of modification may display distinct functions when interacting with different factors (Zentner and Henikoff, 2013). Considering DNA methylation, an overall increase in methylated cytosine is generally a sign of transcriptional suppression. However, it requires a more thorough investigation of methylated CG, CHG and CHH landscape to interpret the epigenetic silencing status of a particular genomic region. For instance, substantial CG methylation has often been found in the gene-body of active genes (Law and Jacobsen, 2010). Small RNAs (sRNAs) comprise several types of RNA fragments ranging from 16 nt to 35 nt with multiple origins and biogenesis pathways that display distinct functions (Borges and Martienssen, 2015; Schorn and Martienssen, 2018). Small RNAs specialize in gene or TE silencing by triggering transcriptional gene silencing (TGS), post-transcriptional gene silencing (PTGS), or inhibition of TE reverse transcription. Following the findings in the previous chapter, we wanted to determine whether interrogating the small RNA components of embryogenic callus could shed some light on the dynamics of epigenetic regulation of TEs under our test conditions. An experiment to determine global methylation as part of this study was planned, but time and financial resources prevent us from presenting this data here.

### 8.2.1 Micro RNA (miRNA) in plants

Micro RNAs originate from endogenous noncoding genes, often designated as *MIR* genes, via an RNA Pol II-dependent mechanism. Generally, these genes are transcribed by Pol II into single-stranded and polyadenylated RNAs, as known as primary miRNAs (pri-miRNAs), which are then partially folded into a stem-loop structure (Borges and Martienssen, 2015). The ribonuclease III Dicer 1 (DCL1) then recognises the stem-loop conformation and process this into a mature 20-22 nt miRNA duplex (Borges and Martienssen, 2015; Cedillo-Jiménez et al., 2016). In plants, miRNAs are a major factor in post-transcriptional gene silencing (PTGS), in which Argonaute proteins (AGOs), AGO1 in most cases, direct miRNAs to the targeted RNA sequences that are highly complementary to the miRNA sequences (Bologna and Voinnet, 2014; Borges and Martienssen, 2015; Cedillo-Jiménez et al., 2016). This interaction usually leads to the cleavage of the targeted RNA molecule; however, in some cases, translation of the target mRNA is inhibited instead.

Most miRNA genes are specific to individual plant families or species, suggesting relatively recent evolution or rapid turnover rate (Borges and Martienssen, 2015; Cuperus et al., 2011). These family- or species-specific miRNA genes are considered to be 'young', whereas a minority of miRNA genes that are highly conserved across various plant families are comparatively 'old' (Cuperus et al., 2011). Both 'young' and 'old' miRNA genes often exhibit tissue and developmentally-specific transcription profiles, with the potential to regulate the transcript abundance of protein-coding genes with respect to the specific stage of plant growth and development (Cuperus et al., 2011). For instance, Lizamore and Winefield (2017) identified several miRNAs that are presented in grapevine leaf and embryogenic callus as some of the miRNAs, e.g. vvi-miR2118 and vvi-miR482, were mutually exclusive in the two tissues. One of the aforementioned 'old' miRNA gene family, *MIR159*, is found to be present in most of the land plants (Cuperus et al., 2011), but the transcriptional profile of the gene members belong to *MIR159* varies in species (Allen et al., 2007; Leng et al., 2015). In *Arabidopsis* and *V. vinifera*, *MIR159* contains three members designated as *MIR159a* to *MIR159c*. In *Arabidopsis*, miR159a and miR159b are plentifully accumulated in the shoot apical region, inflorescences and imbibed seeds, whereas miR159c has low accumulation in the shoot apical region (Allen et al., 2007). In grapevine, vvi-miR159c is generally found to be more abundant than either vvi-miR159a and vvi-miR159bin inflorescences, flowers and embryogenic callus (Leng et al., 2015; Lizamore and Winefield, 2017). Similar to the *Arabidopsis* miR159a/b, grapevine vvi-miR159a/b is also well known for targeting MYB transcriptional factors (Kullan et al., 2015). However, vvi-miR159c is predicted to preferentially target a *Glyoxalase* I protein that potentially participates in stress response (Kullan et al., 2015). These indicate that miRNA members within the same miRNA gene family may be primed for distinct roles.

### 8.2.2 Small interfering RNA (siRNA) in plants

In plants, siRNAs mostly exist as 20-24 nt forms. The smaller 21-22 nt siRNAs are frequently involved in PTGS, while 23-24 nt siRNAs predominantly mediate RdDM (Borges and Martienssen, 2015; Cuerda-Gil and Slotkin, 2016; Fultz et al., 2015). Long double-stranded RNAs (dsRNAs) that give rise to siRNAs can be formed by multiple mechanisms, such as by the hybridization of sense and antisense transcripts, self-complementary binding of two inverted repeats, complementary binding of two RNA molecules derived from distinct loci, or by the activity of RNA-dependent RNA Polymerases (RDRs; Borges and Martienssen, 2015).

Long dsRNAs originated from RNA-Pol II-transcribed RNA molecules are predominantly the precursors of 21-22 nt primary siRNAs processed by DCL1, DCL2 and DCL4. The biogenesis of primary siRNAs is considered as a front line response to TE transcriptional activation. These primary siRNAs are bound within AGO1, which allows targeting of the AGO1-siRNA complex to the mRNAs

complementary to the siRNA sequences and subsequently cleaves mRNAs. The resulting mRNA fragments are then primed by RDR6 for dsRNA synthesis. The dsRNAs produced by RDR6 are trimmed by DCL2 and DCL4 into 21-22 nt in length and serve as secondary siRNAs that amplify the magnitude of the PTGS circuit (Figure 8.1 A; Cuerda-Gil and Slotkin, 2016). A subset of the 21-22 nt secondary siRNAs is manufactured through a phasing procedure that results in uniformly aligned terminus (Borges and Martienssen, 2015; Cedillo-Jiménez et al., 2016). These phased siRNAs are termed tasiRNA, providing the trans-acting property that targets loci distinct from their original ones.

RNA transcripts generated by plant-specific Pol IV can bind with RDR2 to serve as templates for the production of dsRNAs, which are then processed by DCL3 into 24 nt siRNAs before associating with AGO4/6 to then target Pol V-transcribed RNA molecules. The binding of the AGO4/6-siRNA complex and Pol V-transcribed RNA molecules does not lead to the cleavage of the RNA molecule; instead, it facilitates the recruitment of DNA methyltransferase DRM2 to deposit methyl-cytosine on the DNA locus transcribed by Pol V and accomplishes RdDM (Figure 8.1 A; Cuerda-Gil and Slotkin, 2016).

### 8.2.3 Transfer RNA (tRNA) and ribosomal RNA (rRNA) fragments

Small RNA sequencing reads derived from tRNAs and rRNAs are often discarded in sRNA analysis. However, increasing evidence shows that tRNAs and rRNAs can be degraded asymmetrically into size-specific fragments, suggesting specialized roles of these fragments other than products of random degradation (Li et al., 2012b).

In addition to the well-known biological roles of tRNAs in translation, researchers have seen the stress-induced accumulation of size-specific tRNA fragments (tRFs) in animals, plants and yeast, and this is not necessarily associated with impaired tRNA biogenesis (Thompson et al., 2008). Matured tRNAs are characterised with the 3’ CCA tail and three-leafed-clover-like structure that contains three hairpin loops. With cleavage at different loops, tRNA molecules can be processed into 16-35 nt fragments retaining the 5’ end, the intermediate part without both ends, or the 3’ CCA-containing tail (Schorn and Martienssen, 2018). The interaction between tRFs and all four human AGO proteins (AGO1 - 4) implies the regulatory role of tRFs in gene silencing (Haussecker et al., 2010; Kumar et al., 2014). In mouse, a 5’ tRF, tRNA-Gly-GCC fragment, was found capable of repressing genes associated with an endogenous LTR-TE in both mouse embryos and embryonic stem cells (Sharma et al., 2016). Likewise, *Arabidopsis* 5’ tRFs were reported to be processed by DCL1 and interact with AGO1, while a 5’ tRF was shown to bind to complementary sites in gypsy mRNAs and removes these via cleavage within the target binding site (Martinez et al., 2017). On the other hand, 3’ tRFs, especially those of 18 nt and 22nt in length, exhibit inhibitory properties against LTR-TEs (Schorn et al., 2017). The primer binding site (PBS) of LTR retrotransposons is located directly after the 5’ LTR. This site serves

as a binding site for host tRNAs that act to initiate reverse transcription of the transcribed TE-RNA (Schorn and Martienssen, 2018). The PBS sequences were often found led by conserved TGG codon in animal and plants (Borges et al., 2018; Neumann et al., 2019; Schorn and Martienssen, 2018; Schorn et al., 2017). These sequences can act as a target, not only of tRNAs for priming reverse transcription but also of 3' tRFs containing CCA tail (Schorn et al., 2017). In mouse, it has been shown that the 18 nt 3' CCA-tRF can block the priming of mature tRNAs at PBS, thus inhibit reverse transcription, whereas the 22 nt 3'CCA-tRFs are capable of inhibiting translation of TE proteins and triggering post-transcriptional silencing by associating with AGO2 (Li et al., 2012b; Schorn et al., 2017). Whether this silencing mechanism exists in plant system remains unknown.

Unlike tRFs, the role of rRNA fragments in gene silencing has not yet been reported. Digested rRNA fragments 300-600 nt in length have been observed in stress-treated plants and linked with cell death, while different stresses may lead to distinct digestion patterns of these rRNA derived RNA species (Hoat et al., 2006; Mroczek and Kufel, 2008). It has been reported that the abundance of 18-25 nt rRNA fragments in Arabidopsis pollen is 4.14-fold higher than that in *Arabidopsis* inflorescence (Martinez et al., 2017). However, this phenomenon was suggested to be a consequence of increased rRNA transcripts and/or degradation in pollen since the increase of rRNA fragments in pollen appears to be evenly contributed by all sizes of rRNA fragments (Martinez et al., 2017). It is unclear whether the finding regarding rRNA fragments in the comparison of *Arabidopsis* pollen and inflorescence is conserved in grapevine embryogenic callus subjected to HDACi. It is sensible to speculate that an even distribution or a gradual distribution towards the smallest size category might indicate a random degradation of rRNAs (Cholet et al., 2019; Hoat et al., 2006; Martinez et al., 2017). Alternatively, a size distribution clearly peaked at specific categories might imply the existence of specialized functions (Martinez et al., 2017; Schorn et al., 2017).

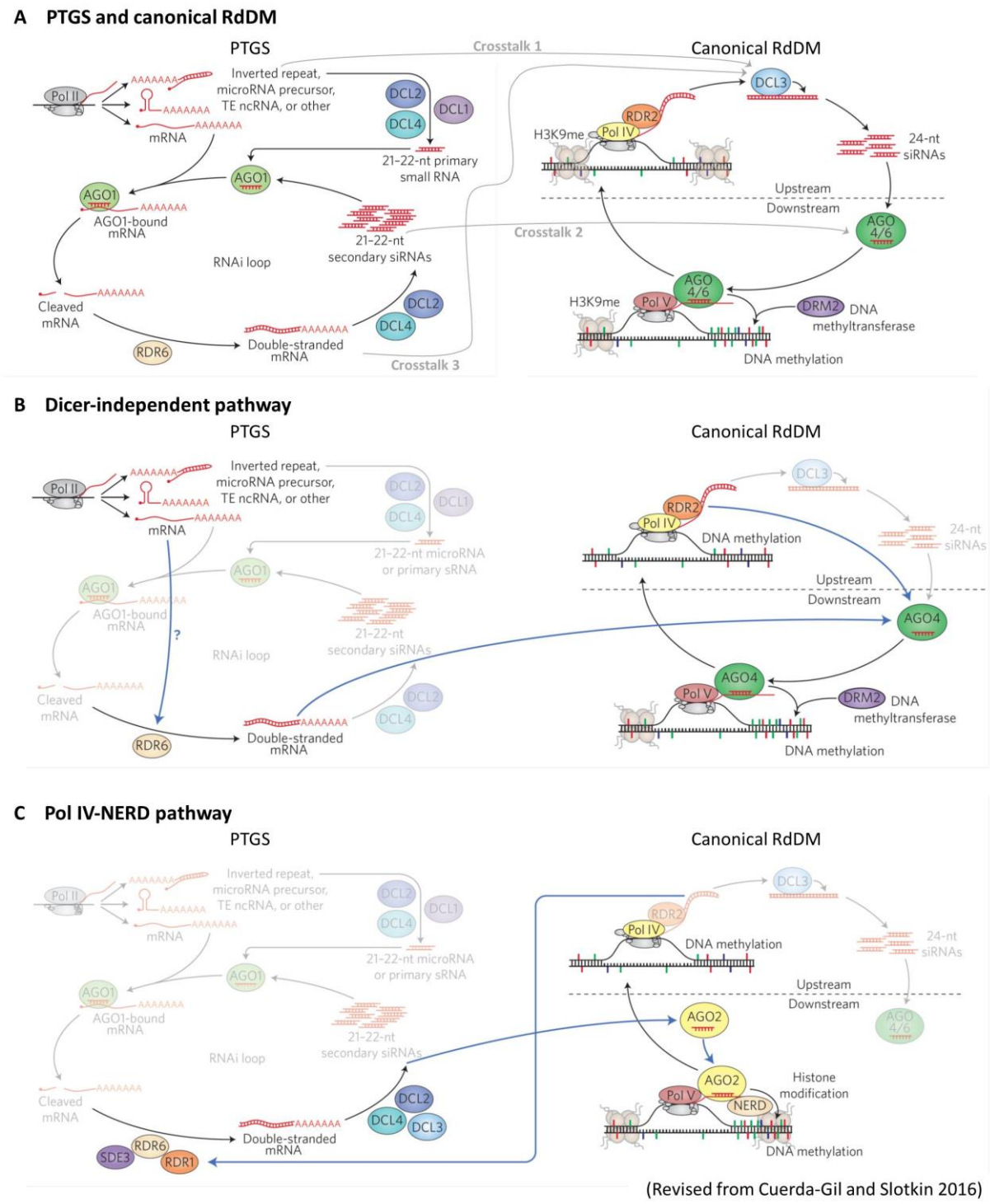



**Figure 8.1 PTGS and the canonical RdDM pathway in plants**

**(A)** In PTGS, Pol II-transcribed RNAs are precursors of 21-22 nt primary siRNA processed by DCL1, DCL2 and DCL4. The deployment of primary siRNA by AGO1 to the complementary site of mRNA triggers cleavage on mRNAs, which is then used as templates for dsRNA synthesis by RDR6. The dsRNAs are processed by DCL2/4 to generate 21-22 nt secondary siRNAs that can bond with AGO1 and thus amplify the magnitude of PTGS. In the upstream of canonical RdDM, loci with heterochromatic hallmarks, e.g. H3K9me, and DNA methylation, are primed by Pol IV and RDR2 for the synthesis of dsRNAs that are further processed by DCL3 to generate 24 nt siRNA. By interacting with AGO4/6, the 24 nt siRNAs act in trans to target distinct heterochromatic loci transcribed by Pol V. This recruits DNA methyltransferase DRM2 to deposit methyl group densely on cytosine. Interconnections between PTGS and RdDM are denoted as Crosstalk 1-3. **(B)** A dicer-independent pathway that RDR6 and AGO4 bridge PTGS and RdDM. **(C)** A Pol IV-NERD pathway that the upstream and downstream of RdDM are connected by various RDR and DCL before feeding siRNAs to AGO2 to trigger histone

modification and DNA methylation in the presence of NERD. Please see 8.2.2 and 8.2.4 for more details. This figure is modified from Cuerda-Gil and Slotkin (2016).

### 8.2.4 Crosstalk between PTGS and RdDM

Various forms of interplay between RdDM and PTGS have been reported in many plant species. In *Arabidopsis*, rice and moss, Pol II-generated pri-miRNAs or dsRNAs can be trimmed by DCL3 into 24 nt siRNAs, therefore enabling RdDM with the presence of AGO4/6, Pol V, and DNA methyltransferase DRM2 (Figure 8.1 A Crosstalk 1; Chellappan et al., 2010; Khraiwesh et al., 2010; Wu et al., 2010). Additionally, the 21-22 nt secondary siRNAs produced by DCL2 and DCL4 can be fed into the RdDM pathway by AGO6 (Figure 8.1 A Crosstalk 2; McCue et al., 2015). Using *Arabidopsis* epigenetic Recombinant Inbred Lines (epiRILs) that allow accumulation of novel TE insertions, Mari-Ordonez and colleagues (2013) demonstrated a mechanism of *de novo* silencing reconstruction, in which RDR6 converted mRNAs of the novel TEs into dsRNAs, following with DCL3-dependent production of 24 nt siRNAs that interact with AGO4, eventually establishing DNA methylation on these novel TEs (Figure 8.1 A Crosstalk 3). Alternatively, PTGS and RdDM can be bridged in a Dicer-independent mechanism, where RDR6-generated dsRNAs can be directly incorporated into AGO4 without processing by Dicer endonuclease; instead, these AGO4-incorporated dsRNAs are subsequently trimmed by 3'-5' exonuclease into proper size for RdDM (Figure 8.1 B; Cuerda-Gil and Slotkin, 2016; Ye et al., 2016). In a less reported non-canonical RdDM pathway (Figure 8.1 C; Cuerda-Gil and Slotkin, 2016), dsRNAs synthesised by Pol IV, RDR1 and RDR6 with the presence of RNA helicase SDE3 can be processed by DCL2, DCL3 and DCL4 into smaller fragments that bind to AGO2 (Cuerda-Gil and Slotkin, 2016). The AGO2-siRNA complex is then navigated to Pol V-transcribed nascent RNAs with the assistance of NERD (Needed for RDR2-independent DNA methylation; Pontier et al., 2012), and eventually result in RdDM (Cuerda-Gil and Slotkin, 2016). Notably, AGO2 has been reported involving in PTGS by interacting with specific miRNA to mediate antibacterial defence in *Arabidopsis* (Zhang et al., 2011), suggestive of its dual roles in PTGS and RdDM.

These various types of crosstalk between PTGS and RdDM might be particularly crucial when confronting extreme TE activation or the presence of novel TEs in the genome. PTGS was reported to be the first-order response when there is an extensive increase in TE activity derived from newly introduced TEs into the host genome (Marí-Ordóñez et al., 2013). During sexual reproduction in *Arabidopsis* pollen, a re-activation of TE transcription follows a down-regulation of chromatin remodeler DDM1 and 24 nt siRNA biogenesis in the vegetative nucleus, which accompanies the sperm cells in mature pollen grains but does not transmit DNA to the fertilized zygote (Slotkin et al., 2009). In turn, 21 nt siRNA specialised in PTGS was accumulated in the vegetative nucleus and introduced to the sperm nuclei for the establishment of methylation patterns in gametes that

prevent significant TE activity. In heat-shocked *Arabidopsis* wild-type, heat-activated retrotransposon *ONSEN* exhibits increased *ONSEN* transcripts and transposes at low frequency (Ito et al., 2011). However, in heat-shocked *Arabidopsis* mutant *nrpd1*, which is deficient in a Pol IV subunit that is required for RdDM pathway, the transcriptional level of *ONSEN* was significantly higher than that in heat-treated wild-type, and the copy number of new *ONSEN* insertions is 8-22 times as many as that in wild-type background depending on the duration of heat treatment and the recovery phase after heat shock (Ito et al., 2011). Accumulation of 21 nt siRNAs was found in both wild-type and *nrpd1* subjected to heat shock. Although the level of 21 nt siRNA in *nrpd1* is significantly higher than that in wild-type, this siRNA is not able to prevent the accumulation of *ONSEN* transcripts in *nrpd1* (Ito et al., 2011). These examples suggest that RdDM is the main mechanism targeting pre-existing genomic TEs, whereas PTGS serves as a second mechanism coming to the fore when RdDM is compromised or unable to recognize novel TEs.

### 8.2.5 Scenarios depicting the balance between sRNA and TE activity

The current understanding of silencing pathways in plants has been largely established using epigenetically compromised mutants, in which key factors in PTGS or RdDM, such as chromatin remodeler DDM1, DNA methyltransferase MET1, Argonautes (AGOs), Dicers (DCLs) or RDR families, were depleted. Through investigation of TE activity and the dynamic epigenetic landscape, including DNA methylation and histone marks (e.g. H3K9 and H3K27 methylation as well as H3K9 and H4K16 acetylation), researchers have harnessed these mutants to reconstruct the interwoven networks node by node. However, it is still unclear about how, in the wild-type backgrounds, these networks coordinate together to respond to inevitable TE transcriptional perturbation or how these pathways are able to allow low-frequency accumulation of new TE insertions that have been observed in various plant species (Baduel et al., 2019; Hashida et al., 2006; Lizamore, 2013; Rakocevic et al., 2009). In most of the cases, the accumulation of TE transcripts was presented by northern blot, qPCR or short-read-based sequencing technology. However, few TE insertion was introduced unless the stressed plants were deficient in part of the silencing machinery such as DNA methylation or siRNA synthesis (Ito et al., 2011). In our research, wound-like treatment, biotic stress, and pharmacological inhibition of HDAC led to the TE transcriptional perturbation, as revealed by the Illumina RNAseq. However, competent transcripts from autonomous TEs were undetectable in the ONT cDNA sequencing data. Therefore, it is proposed that proper stress treatment customized to the stress-related CREs in TEs, as well as a compromised epigenetic system, are both required to trigger large-scale autonomous TE mobilization (see chapter 7). In other words, it is likely that, without either of these two factors, the epigenetic system, particularly PTGS, would still be competent in neutralizing

TE activity in a dosage-dependant manner, unless the stimulated TE activity is high enough to saturate the epigenetic machinery.

The *ONSEN* retrotransposon is well-known for its heat responsiveness in rice and *Arabidopsis*. Using a more sensitive method to detect new TE insertions, Thieme et al. (2017) showed that heat shock did increase the copy number of *ONSEN* in wild-type *Arabidopsis* seedlings, whereas drug inhibition of RdDM pathway in addition to heat shock treatment further increased the magnitude of TE transposition. As elevated levels of TE activity are frequently accompanied by an increase of miRNA or 21-22 nt siRNA (Borges et al., 2018; Ito et al., 2011; Marí-Ordóñez et al., 2013; Slotkin et al., 2009), there is a possibility that wild-type plants respond to increased TE transcripts with elevated miRNA or siRNA levels targeting these TEs. The miRNA/siRNA abundance is likely to be dominant over the stress-induced increase in TE activity unless the level of new TE transcription saturates the silencing system (Figure 8.2 A). However, if the silencing machinery is compromised, the relative level between regulatory sRNAs and TE transcripts may be reversed as the epigenetic factors available to silence TE transcription are depleted (Figure 8.2 B). The early stage of Figure 8.2 A may then explain the situation in the stress- or drug-treated grapevine embryogenic callus, where the dominant activity of inhibitory sRNAs is synchronized with TE dynamics (Figure 8.2 C). Alternatively, the inhibitory sRNAs might be maintained at a sufficient and steady level irrespective of TE dynamics as long as the dosage of new TE transcripts fails to saturate the system to the trigger point, where the epigenetic machinery would sense and act to accelerate the accumulation of sRNAs (Figure 8.2 D).

To check whether the epigenetic machinery, in terms of sRNA activity, was competent to deal with grapevine embryogenic callus treated with HDACi, and to understand how host cells harness sRNAs in response to TE perturbation, small RNA samples isolated from the exactly same callus used in chapter 7 were sequenced and analysed. Combining the Illumina RNAseq data presented in chapter 7, this chapter will discuss the possible scenario of how epigenetic networks cope with TE perturbation in grapevine embryogenic callus.

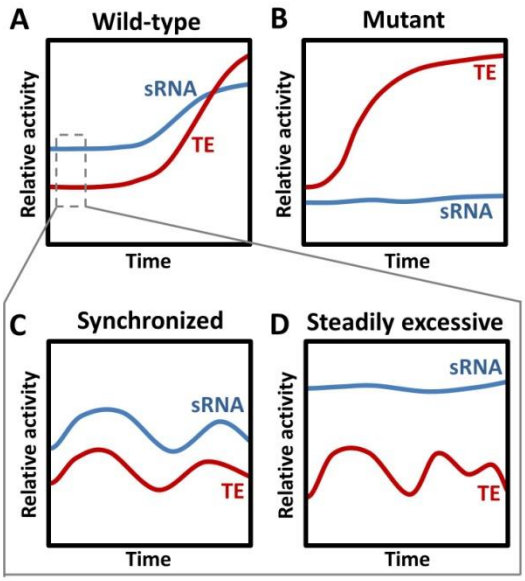



**Figure 8.2 Proposed models of the balance between regulatory sRNAs and TE activity.**

**(A)** In a wild-type background, extreme environmental cue (e.g. heat) may activate TE transcription (e.g. *ONSEN*). In a dosage-dependent manner, the accumulation of sRNAs is accelerated. TE transposition may take place when the level of TE transcripts catches up sRNA level. **(B)** Mutant deficient in epigenetic silencing display unleashed TE activity, especially when stress-stimulated since sRNAs are almost depleted. **(C)** In the early stage of A, the sRNA level might fluctuate with TE dynamics in a synchronized manner, although time lag might happen. **(D)** Alternatively, the TE perturbation is at a level that is far less than the excessive and steady level of sRNAs.

## 8.3 Methods

### 8.3.1 Sample preparation and sequencing

Small RNAs were purified from the treated embryogenic callus samples used in chapter 7 using the NORGEN Plant microRNA purification kit (Norgen Biotek). Small RNA sequencing libraries were prepared using the NEXTflex Small RNA-seq Kit (Bioo Scientific), according to the manufacturer's instructions. In brief, 200 ng of small RNA was initially used for the ligation reaction with 3' 4N-adenylated adapters that specifically ligated to the 3' hydroxyl groups of sRNAs. After removal and inactivation of excess 3' adapters, the resulting products were ligated with 5' 4N adapters at the 5'-phosphate group, and then samples were subjected to reverse transcription for the synthesis of the first-strand cDNA. Reaction products were then subjected to 9 to 10 cycles of PCR amplification before bead-based size selection and clean-up. The libraries were quantified using Qubit dsDNA HS Assay Kit (Thermo Fisher), and quality checked with Bioanalyzer High Sensitivity DNA Assay (Agilent) to make sure the presence of a single clear peak with the size around 150-152 bp. The resulting stranded libraries were then sent to GENEWIZ (https://www.genewiz.com/) to be sequenced (single-end) on an Illumina HiSeq X-TEN platform.

### 8.3.2 Pre-processing of sequencing data

According to the recommendation from the manufacturer's manual, the 3' adapter sequence was removed from the sequencing reads following with trimming of the first and last four bases from the adapter-clipped reads, since the 3' and 5' adapters of NEXTflex Small RNA-seq Kit contain four random bases to reduce sequence bias in preparing sRNA sequencing library. Sequencing reads that were 16 nt to 35 nt in length were then selected and aligned to the *Vitis vinifera* reference genome using Bowtie (Langmead, 2010; Langmead et al., 2009). All the alignment procedures described in this chapter were conducted using Bowtie with the parameters `-v 0 -k 100 -a --best --strata`, which allow the output of multiple alignments up to 100 valid alignments per read and only retaining perfect-match reads for downstream analysis.

Each category of sRNA was quantified non-redundantly (i.e. ignoring multiple alignments) and normalized as reads per million total mapped (RPM). Due to the scale of the experiments having five time points (0, 12, 24, 48 and 72 hours), each with three technical replicates, mock, TSA and 4PBA treatments were conducted in three different but consecutive days. To establish a comparable time zero (T=0) across all treatments, the sRNA accumulation levels at treated time-points (12, 24, 48 and 72 hours) were divided by the corresponding T=0. In this way, the sRNA relative accumulation at T=0 would be normalized as 1.

### 8.3.3 tRNA and rRNA fragment analysis

Reads perfectly mapping to the reference list of grapevine tRNA sequences (Repbase) were collected for tRNA fragment (tRF) analysis. Following the protocol of Schorn et al. (2017), reads were grouped into 3' CCA- and 3' non-CCA-ending fragments and sorted by size. The rest of the unmapped reads were then aligned to the collection of grapevine rRNA sequences (Repbase) and sorted by size.

### 8.3.4 miRNA analysis

Following the steps in 8.3.3, reads not mapping to tRNA and rRNA were aligned to the list of miRNA sequences established by Lizamore and Winefield (2017). The alignment suppressed mapping against the reverse-compliment reference strand by specifying `--norc` in Bowtie. Multi-mapping reads were assigned randomly to a single locus by Bowtie with the parameter `-M 1`. Except for calculating RPM and relative accumulation of miRNAs, the number of reads mapping to each locus were calculated by bedtools coverage. These raw counts were used in differential analysis with the software DESeq2, which further logarithmically transformed the raw counts into variance stabilizing transformations (VST). miRNA loci with adjusted p-value < 0.05 and VST fold-change > 1 were considered differentially expressed. To predict miRNA targets, miRNA sequences were compared with the canonical set of TE sequences obtained by Lizamore (2013) and the *Vitis vinifera* CRIBI V2.1 transcript sequences using psRNA-Target (Dai and Zhao, 2011). A stringent level of expectation for complementarity (E≤2) was used to collect TEs and genes that are highly likely to be targeted by the miRNA sequences.

### 8.3.5 siRNA analysis

Continuing from 8.3.4, the rest of the unmapped reads were re-aligned to the grapevine reference genome using Bowtie settings as same as in miRNA analysis. Perfectly aligned reads were calculated, normalized to RPM or T=0, and sorted by size. To capture TE-derived siRNA, these siRNA reads were mapped to the canonical TE sequences with the same Bowtie settings, in which multi-mapping reads were randomly assigned to a perfectly matched site. Reads derived from each TE family were calculated using bedtools coverage. The identification of differentially accumulated TE-derived siRNA was conducted in the same way as in 8.3.4.

### 8.3.6 tasiRNA analysis

Among the siRNA reads, a subset of them was processed in phase and termed tasiRNA (see 8.2.3). The total siRNA reads collected in 8.3.5 were analysed by the UEA Small RNA Workbench (Stocks et al. 2018) to identify tasiRNAs and tasiRNA loci. The estimated genic coordinates of tasiRNA loci may be slightly different across treatments. Those overlapped were combined using bedtools merge to generate a conserved list of expressed tasiRNA loci across all treatments. In addition, this combined

set of tasiRNA loci were examined for the intersection with all annotated TEs using bedtools intersect. To quantify the productivity of each tasiRNA loci, tasiRNA reads with multiple hits were divided uniformly to each locus before summing up with uniquely mapped reads. The total count of each locus was then rounded up to the closest integer before being fed into DESeq2 for differential analysis. The criteria for differential expression were as same as in 8.3.4.

The computational scripts can be found in Appendix D.5.

## 8.4 Results

### 8.4.1 Alignment statistics

The sRNA sequencing generated 7 to 16 million sequencing reads. About 67% to 93% of sequenced reads were kept after adaptor removal, most of which were retained after selection for sizes between 16 and 35 nt (Table 7.1). After quality filtering, there were 3.5 to 7.8 million reads mapped to the *V. vinifera* reference genome.

**Table 7.1. Mapping statistics for Illumina Truseq RNA-seq.**

| Sequenced libraries | | | Sequenced reads | | Adaptor removal | | Size selected (16-35 nt) | | Mapped reads | |
|---|---|---|---|---|---|---|---|---|---|---|
| Treatments | Timepoints | Replicates | | | | | | | | |
| Mock | 00 h | a | 7,482,659 | 100% | 6,883,759 | 92.00% | 5,971,889 | 79.81% | 4,207,586 | 56.23% |
| | | b | 8,935,499 | 100% | 8,241,545 | 92.23% | 7,264,199 | 81.30% | 5,170,091 | 57.86% |
| | | c | 9,311,225 | 100% | 8,672,951 | 93.15% | 7,472,919 | 80.26% | 5,311,059 | 57.04% |
| | 12 h | a | 8,441,056 | 100% | 7,357,145 | 87.16% | 6,239,887 | 73.92% | 4,472,815 | 52.99% |
| | | b | 8,751,700 | 100% | 7,806,024 | 89.19% | 6,552,182 | 74.87% | 4,688,109 | 53.57% |
| | | c | 9,616,192 | 100% | 8,135,529 | 84.60% | 6,797,065 | 70.68% | 4,833,850 | 50.27% |
| | 24 h | a | 7,773,061 | 100% | 6,597,498 | 84.88% | 5,474,215 | 70.43% | 3,902,350 | 50.20% |
| | | b | 9,711,524 | 100% | 8,175,492 | 84.18% | 6,844,961 | 70.48% | 4,908,258 | 50.54% |
| | | c | 7,708,533 | 100% | 6,659,209 | 86.39% | 5,424,305 | 70.37% | 3,870,392 | 50.21% |
| | 48 h | a | 7,172,333 | 100% | 6,109,504 | 85.18% | 5,051,153 | 70.43% | 3,518,657 | 49.06% |
| | | b | 8,820,211 | 100% | 7,682,257 | 87.10% | 6,334,264 | 71.82% | 4,426,384 | 50.18% |
| | | c | 7,823,525 | 100% | 6,766,633 | 86.49% | 5,680,989 | 72.61% | 3,971,369 | 50.76% |
| | 72 h | a | 9,146,297 | 100% | 7,710,210 | 84.30% | 6,335,838 | 69.27% | 4,346,832 | 47.53% |
| | | b | 10,075,196 | 100% | 8,854,835 | 87.89% | 7,149,899 | 70.97% | 4,908,139 | 48.72% |
| | | c | 10,400,405 | 100% | 9,026,255 | 86.79% | 7,492,635 | 72.04% | 5,134,216 | 49.37% |
| TSA | 00 h | a | 9,158,741 | 100% | 7,853,896 | 85.75% | 6,740,727 | 73.60% | 4,577,487 | 49.98% |
| | | b | 10,301,138 | 100% | 9,133,971 | 88.67% | 7,849,005 | 76.20% | 5,411,990 | 52.54% |
| | | c | 9,897,501 | 100% | 8,435,811 | 85.23% | 6,914,279 | 69.86% | 4,615,636 | 46.63% |
| | 12 h | a | 9,861,292 | 100% | 8,034,089 | 81.47% | 6,576,451 | 66.69% | 4,663,070 | 47.29% |
| | | b | 9,773,247 | 100% | 7,926,287 | 81.10% | 6,319,365 | 64.66% | 4,490,727 | 45.95% |
| | | c | 12,674,095 | 100% | 9,696,142 | 76.50% | 8,208,767 | 64.77% | 5,879,070 | 46.39% |
| | 24 h | a | 11,270,054 | 100% | 9,526,044 | 84.53% | 7,717,199 | 68.48% | 5,384,790 | 47.78% |
| | | b | 11,914,004 | 100% | 9,225,649 | 77.44% | 7,587,844 | 63.69% | 5,335,686 | 44.78% |
| | | c | 12,161,759 | 100% | 9,053,418 | 74.44% | 7,510,712 | 61.76% | 5,242,255 | 43.10% |
| | 48 h | a | 11,635,026 | 100% | 9,801,224 | 84.24% | 8,120,052 | 69.79% | 5,512,934 | 47.38% |
| | | b | 9,917,571 | 100% | 8,255,984 | 83.25% | 6,526,032 | 65.80% | 4,437,448 | 44.74% |
| | | c | 11,011,339 | 100% | 7,890,998 | 71.66% | 6,424,232 | 58.34% | 4,442,810 | 40.35% |
| | 72 h | a | 16,746,456 | 100% | 11,290,755 | 67.42% | 8,974,039 | 53.59% | 5,833,896 | 34.84% |
| | | b | 8,833,620 | 100% | 7,184,421 | 81.33% | 5,612,493 | 63.54% | 3,611,507 | 40.88% |
| | | c | 16,760,615 | 100% | 12,535,150 | 74.79% | 10,064,532 | 60.05% | 6,416,132 | 38.28% |
| 4PBA | 00 h | a | 9,817,768 | 100% | 8,858,871 | 90.23% | 7,629,341 | 77.71% | 5,312,713 | 54.11% |
| | | b | 16,356,395 | 100% | 13,083,054 | 79.99% | 11,498,298 | 70.30% | 7,809,523 | 47.75% |
| | | c | 16,059,271 | 100% | 12,808,652 | 79.76% | 11,211,451 | 69.81% | 7,665,748 | 47.73% |
| | 12 h | a | 13,873,270 | 100% | 11,788,188 | 84.97% | 10,294,105 | 74.20% | 7,286,106 | 52.52% |
| | | b | 14,493,071 | 100% | 11,968,330 | 82.58% | 10,302,949 | 71.09% | 7,365,208 | 50.82% |
| | | c | 12,601,650 | 100% | 10,845,159 | 86.06% | 9,194,000 | 72.96% | 6,578,802 | 52.21% |
| | 24 h | a | 12,513,246 | 100% | 10,691,161 | 85.44% | 9,154,485 | 73.16% | 6,515,331 | 52.07% |
| | | b | 11,784,616 | 100% | 10,167,612 | 86.28% | 8,345,284 | 70.82% | 5,859,252 | 49.72% |
| | | c | 11,411,912 | 100% | 10,225,052 | 89.60% | 8,317,448 | 72.88% | 5,842,464 | 51.20% |
| | 48 h | a | 10,325,886 | 100% | 9,210,573 | 89.20% | 7,253,487 | 70.25% | 4,946,990 | 47.91% |
| | | b | 10,660,841 | 100% | 9,357,493 | 87.77% | 7,091,161 | 66.52% | 4,917,146 | 46.12% |
| | | c | 11,032,662 | 100% | 9,859,545 | 89.37% | 7,477,355 | 67.77% | 5,134,988 | 46.54% |
| | 72 h | a | 13,266,197 | 100% | 11,732,347 | 88.44% | 8,229,263 | 62.03% | 5,541,840 | 41.77% |
| | | b | 11,655,471 | 100% | 9,836,947 | 84.40% | 7,442,956 | 63.86% | 4,943,974 | 42.42% |
| | | C | 13,348,565 | 100% | 11,281,253 | 84.51% | 9,179,515 | 68.77% | 6,152,361 | 46.09% |

### 8.4.2 General statistics of sRNA categories

The mapped sRNA reads can be derived from four major groups, including tRNA, rRNA, miRNA and siRNA. As shown in Figure 8.3, siRNAs comprised the majority of the sequenced sRNAs irrespective of treatments, albeit a slight decrease of siRNA proportion over time. Reads derived from tRNAs comprised 2% to 5% of the sRNA pool at T=0, following by 6% to 10% elevations relative to the total pool with the onset of stress treatments. Interestingly, the proportion of reads derived from rRNAs was increased in mock and TSA treatments over time but slightly decreased in the 12 hours and 24 hours of 4PBA treatments. Comparing with the dynamics of siRNA, tRNA and rRNA proportions in the sRNA pool, the proportion of miRNAs was relatively stable across time.

As mentioned previously, these four general sRNA groups contain several types of sRNAs exhibiting various characteristics and functions in epigenetic silencing. The 16 to 35nt tRNA reads are frequently referred to as tRFs, which include those derived from the 3' end of the mature tRNAs characterized with 3' CCA and those without 3' CCA tail. The 21 to 24 nt siRNAs includes those typically derived from heterochromatin (23–24 nt), those directly processed from Pol II-generated transcripts (21-22nt primary siRNAs), and those phased 21-22 nt secondary siRNAs (tasiRNAs). These sRNA species were further investigated.

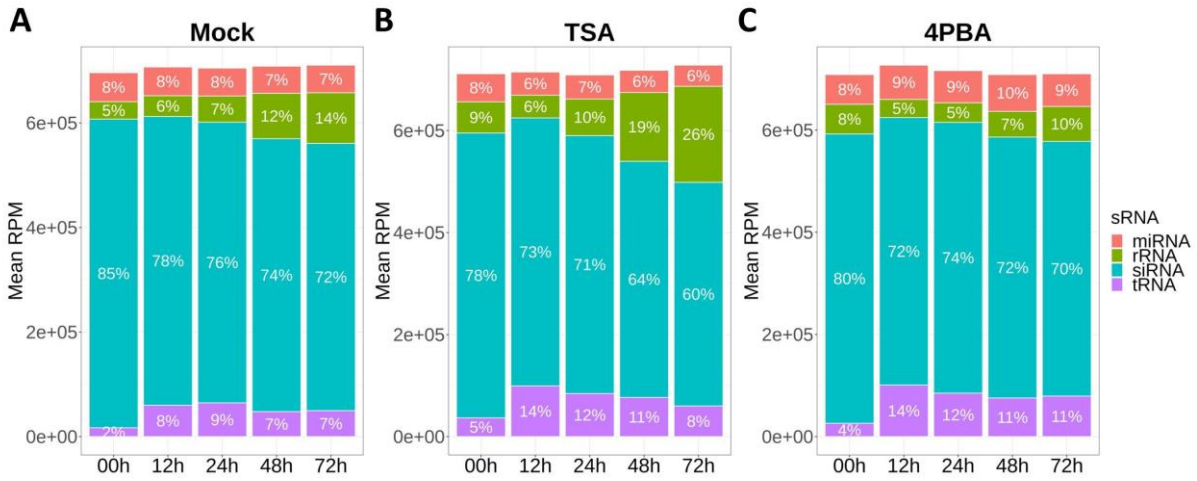



**Figure 8.3 Proportions of sRNA reads derived from tRNA, rRNA, miRNA and siRNA genes**

The mean of total mapped sRNA reads was calculated from the three technical replicates of each time-point and plotted as reads per million total mapped reads (RPM) as shown at the y-axis. These reads were categorised into tRNA, rRNA, miRNA and siRNA groups with their proportion relative to the total mapped read indicated.

### 8.4.2.1 tRNA fragment

As indicated in Figure 8.4 A, there were 20,000 to 30,000 RPM of tRFs at T=0. The amount of tRFs increased with the onset of mock pre-treatment, which was as same as the mock procedure described in section 2.3.1, and continuous incubation with TSA and 4PBA. With normalization to the T=0 samples collected on the same day of the corresponding treatment, the amount of tRFs was elevated at least three-fold higher than T=0 in response to the wound-like pre-treatment (Figure 8.4 B). It peaked in 24 hours (3-4 fold increase) of post-treatment and sustained throughout the sampling time-points. This response pattern and level were not changed in the 4PBA treatment. In TSA treatment, there's also a stimulation of tRFs, yet with a smaller scale, which is significantly milder than the mock response. Grouped by size, the majority of tRFs in grapevine embryogenic callus are 16 nt in length (Figure 8.5). The amount of each 17-19 nt tRFs was about one-third of the 16 nt tRFs, while the population between 20-27 nt was rare. tRFs sizing from 28 nt to 35 nt show a bell-like size distribution that peaked at 32 nt with ~5,000 RPM.

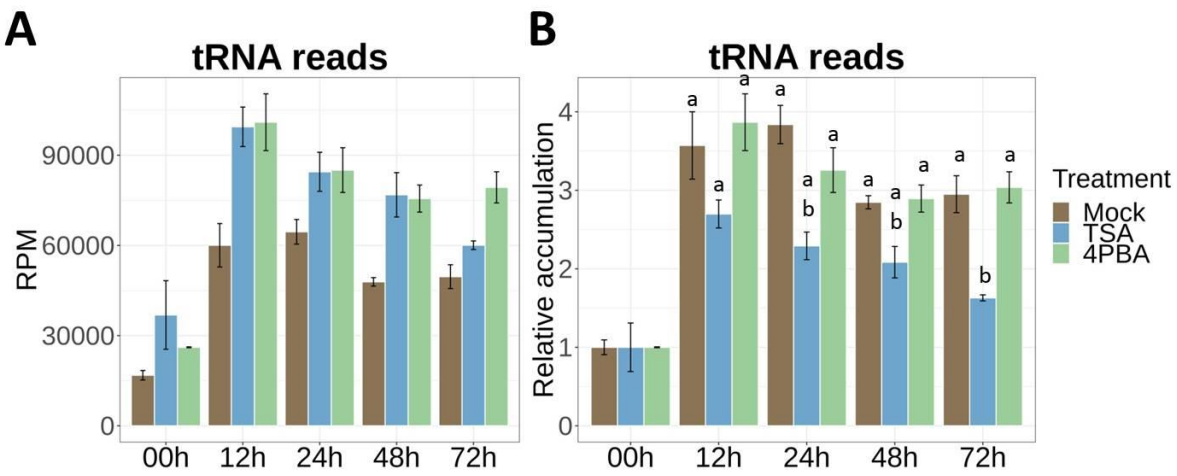



**Figure 8.4 The amount of tRF in multiple treatments over time**

**(A)** The raw counts of tRF were normalized as RPM. **(B)** The relative accumulation of tRF was calculated based on the RPM value relative to T=0. The mean of three replicates ± SD was plotted. The t-test was performed for the relative accumulation, where the comparisons against T=0 (00h) with p-value < 0.05 are denoted by 'a' and the tests with p-value < 0.05 against mock at each time-point are denoted by 'b'. The mock, TSA and 4PBA treatments are shown in brown, blue and green colours, respectively. Time-points were as indicated at the x-axis.

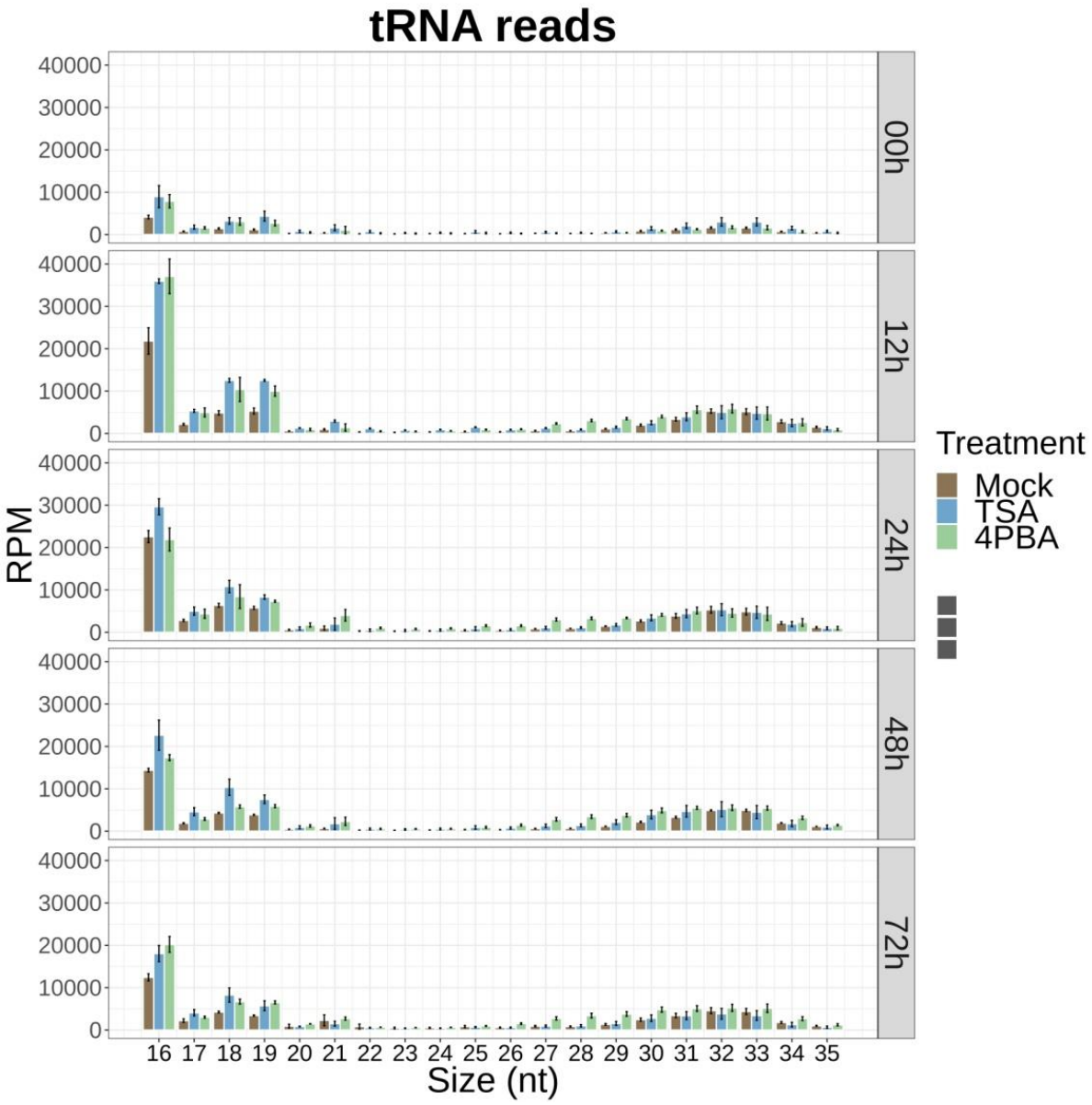



**Figure 8.5 The amount of tRF grouped by size**

The raw counts of tRF grouped by size (x-axis) were normalized as RPM (y-axis). The mean of three replicates ± SD was plotted. The mock, TSA and 4PBA treatments are shown in brown, blue and green colours, respectively. Time-point was as indicated on the right side of each panel.

tRFs can generally be separated into two groups, those containing CCA at 3' end (3' CCA tRFs) and those lacking CCA at the 3' end (3' non-CCA tRFs). Ranging from 50 to 175 RPM, the 3' CCA tRF in mock was about two-fold higher from 12 hours to 48 hours and 3.5 fold higher at 72 hours of post-wound-like-treatment than the initial status (Figure 8.6 A-B). Fluctuation in the relative accumulation of 3' non-CCA tRFs was also seen in TSA and 4PBA treatment. However, the variation of these changes was high among technical replicates, implicating that there's no considerable change in the 3' CCA tRF level. In contrast, the 3' non-CCA tRF was significantly increased in all three treatments

(Figure 8.6 C-D). Because 3’ non-CCA tRFs comprised over 99% of tRFs, their changes basically reflected the changes seen in total tRFs (Figure 8.4).

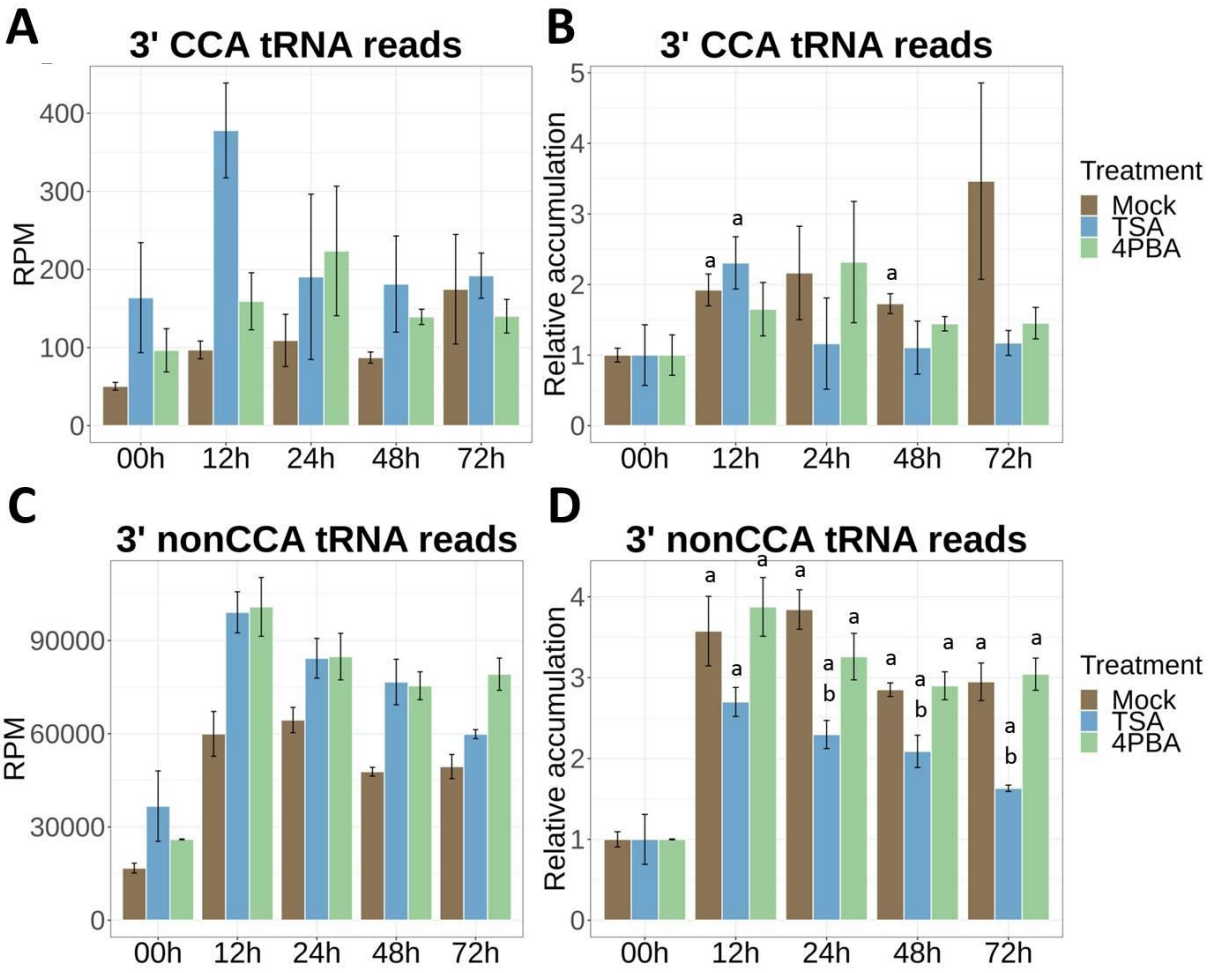



**Figure 8.6 The amount of 3’ CCA and 3’ non-CCA tRF in multiple treatments over time**

**(A)** The accumulation of 3’ CCA tRF is shown as RPM. **(B)** The relative accumulation of 3’ CCA tRF. **(C)** The accumulation of 3’ non-CCA tRF shown as RPM. **(D)** The relative accumulation of 3’ non-CCA tRF. All data were shown as the mean of the three replicates ± SD. The t-test was performed for the relative accumulation, where the comparisons versus T=0 (00h) with p-value < 0.05 are denoted by ‘a’ and the tests with p-value < 0.05 against mock at each time point are denoted by ‘b’.

The 3’ CCA tRFs of grapevine embryogenic callus was enriched in 16-21 nt and 26-29 nt, while the standard deviation of the three replicates was high (Figure 8.7). As mentioned in 8.2, human and mouse 3’ CCA tRFs that are 17 nt to 22 nt in length have been proved to mediate PTGS and inhibit reverse transcription (Schorn and Martienssen, 2018). In our case, the accumulation of 16-21nt 3’ CCA tRFs relative to T=0 was increased in 72 hours of mock and 12 hours of TSA treatment (Figure 8.8). However, the differences of the three technical replicates were too diverse to draw a conclusion.

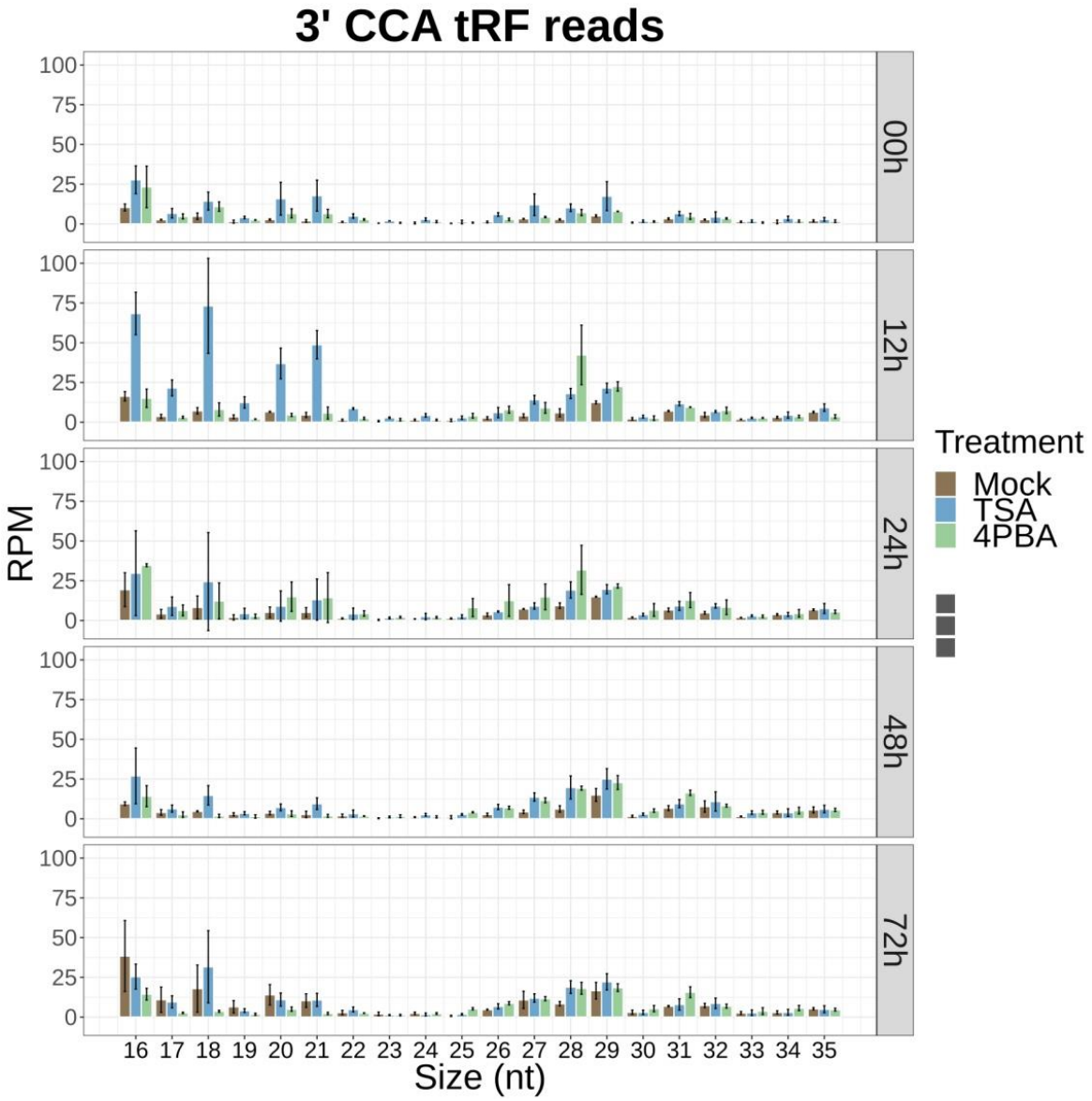



**Figure 8.7 The amount of 3' CCA tRF grouped by size**

3' CCA tRF were grouped by size (x-axis), and the accumulation was shown as RPM (y-axis). The mean of three replicates ± SD was plotted. The mock, TSA and 4PBA treatments are shown in brown, blue and green colours, respectively. Time-point was as indicated on the right side of each panel.

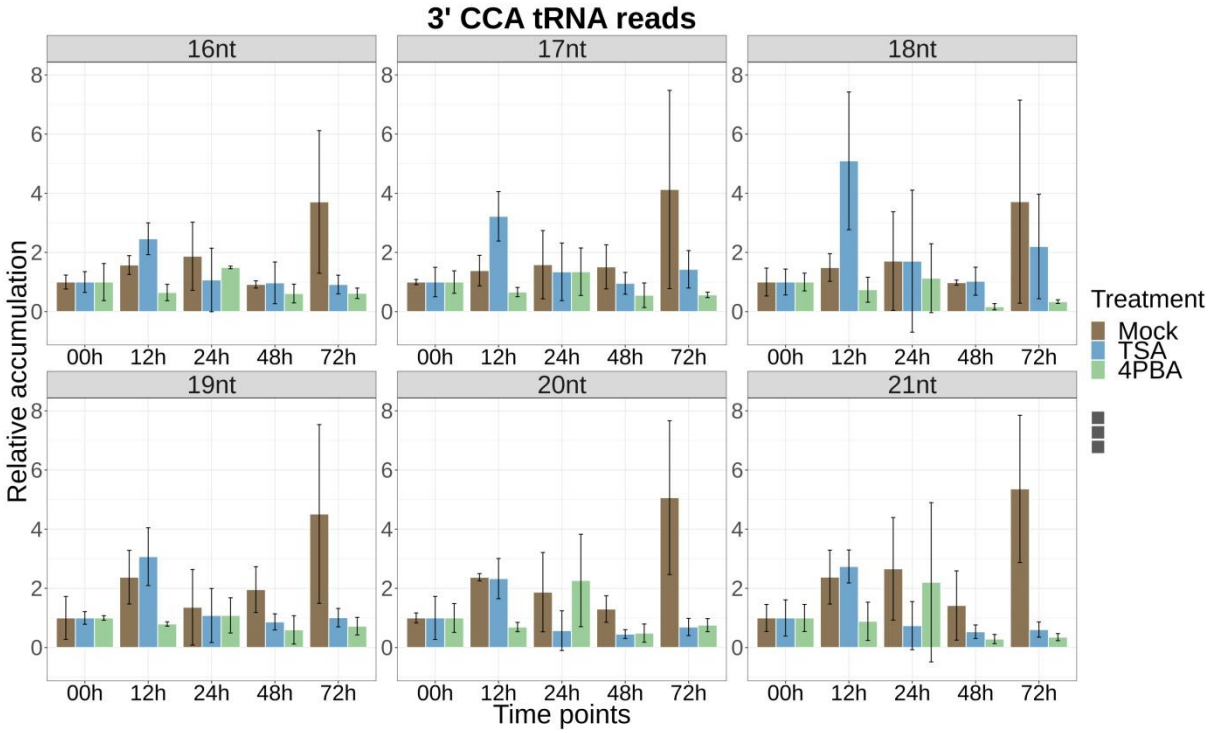



**Figure 8.8 Relative accumulation of 16-21 nt 3' CCA tRF**

The 3' CCA tRF RPM level of each size category was compared with the corresponding level at T=0. The relative accumulation (y-axis) was plotted against the time point (x-axis). All data were shown as the mean of the three replicates ± SD.

In mammals, 3' non-CCA tRFs may include internal fragment (~30 nt), 5' halves (~35 nt), and 5' tRF (19-34 nt), which exhibit various affinities with Argonaute (like AGO and PIWI), cytochrome c (CYC), and Y-box Binding Protein *1* (YBX-1) involving in translational repression (Schorn and Martienssen, 2018). In *Arabidopsis*, 19 nt 5' tRFs were identified with TE-targeting behaviour resembling miRNAs (Martinez et al., 2017). In the grapevine embryogenic callus, the 3' non-CCA tRFs were mostly 16-19 nt and 28-35 nt in length (Figure 8.9). The relative amount of 16 nt to 21 nt 3' non-CCA tRFs revealed elevated accumulation in mock treatment (Figure 8.10), particularly 16-19 nt 3' non-CCA tRFs that resemble the pattern shown in Figure 8.6 D. TSA and 4PBA treatments also increased the level of 16-19 nt 3' non-CCA tRFs, while the fluctuation of 20-21 nt 3' non-CCA tRFs was more diverse among three replicates. However, TSA and 4PBA treatments didn't further enhance tRF accumulation seen in mock. In fact, the increase of 16 nt and 19nt 3' non-CCA tRF in the TSA treatment seems to be lower than the level displayed in mock as of 24 hours of incubation.

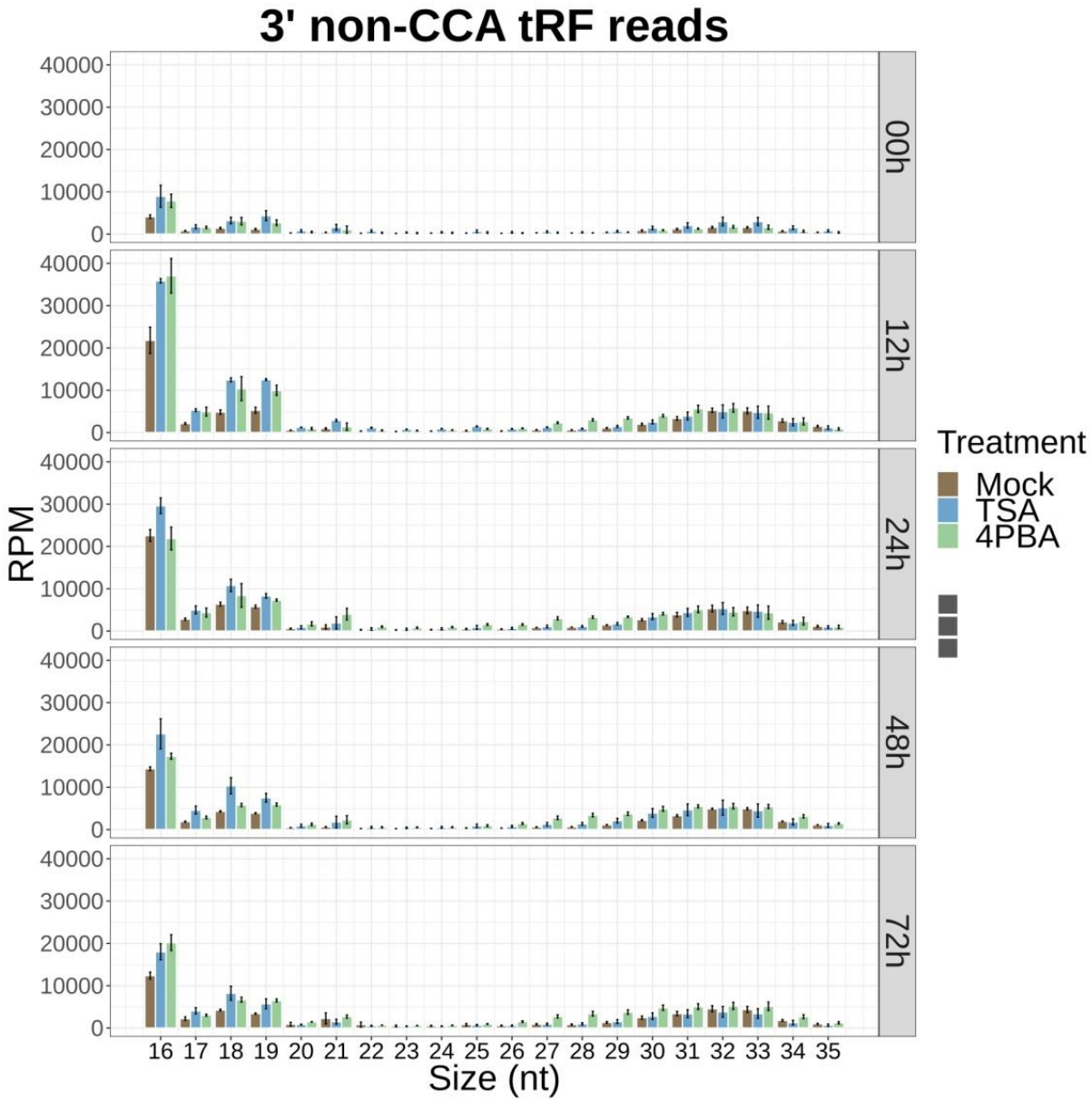



**Figure 8.9 The amount of 3' non-CCA tRF grouped by size**

3' non-CCA tRF were grouped by size (x-axis), and the accumulation was shown as RPM (y-axis). The mean of three replicates ± SD was plotted. The mock, TSA and 4PBA treatments are shown in brown, blue and green colours, respectively. Time-point was as indicated on the right side of each panel.

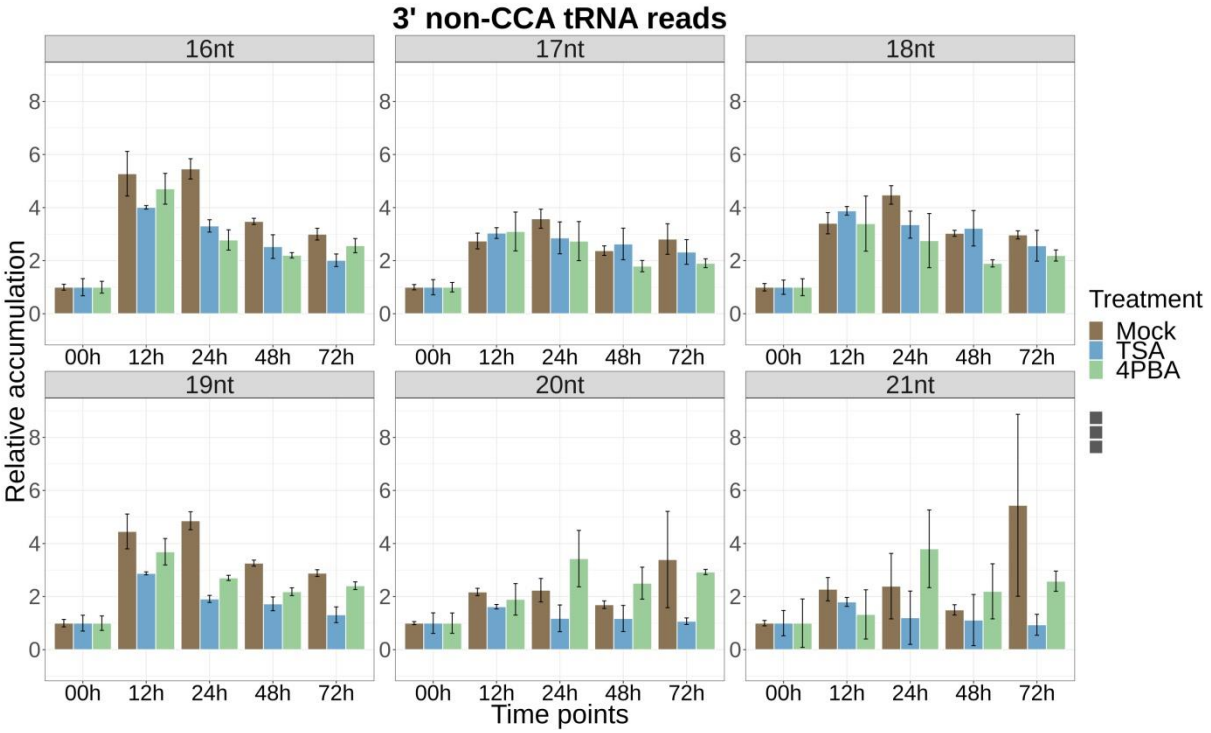



**Figure 8.10 Relative accumulation of 16-21 nt 3' non-CCA tRF**

The 3' non-CCA tRF RPM level of each size category was compared with the corresponding level at T=0. The relative accumulation (y-axis) was plotted against the time-point (x-axis). All data were shown as the mean of the three replicates ± SD.

### 8.4.2.2 rRNA fragment

rRNA fragments were increased in mock treatment over time (Figure 8.11 A-B). The presence of TSA resulted in a similar dynamic accumulation of the rRNA fragment, given that the relative level of 12 hours and 24 hours of TSA treatment was lower than that of the corresponding mock time-point. However, somewhat surprisingly, the level of rRNA reads was not increased in the 4PBA treatment; rather, it decreases at 12 and 24 hours compared to T=0 and then returned gradually back to the initial level.

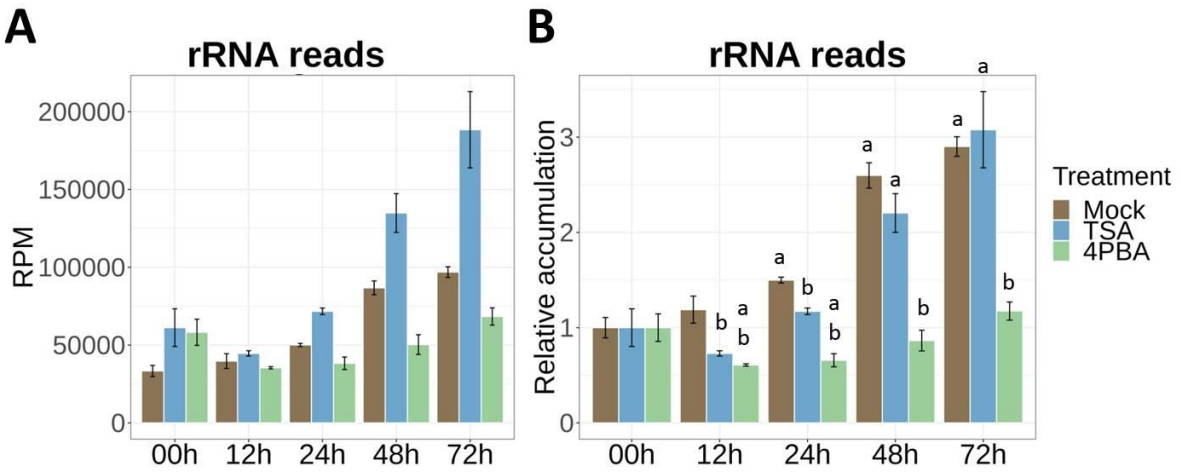



**Figure 8.11 The amount of rRNA fragments in multiple treatments over time**

**(A)** The accumulation of rRNA fragments shown as RPM. **(B)** The relative accumulation of rRNA fragments. All data were shown as the mean of the three replicates ± SD. The t-test was performed for the relative accumulation, where the comparisons versus T=0 (00h) with p-value < 0.05 are denoted by 'a' and the tests with p-value < 0.05 against mock at each time-point are denoted by 'b'.

Analysis of the rRNA fragment size distribution reveals the gradual increase towards the shortest size in all treatment across time (Figure 8.12). This resembles a common pattern of random degradation. Further investigation of 16-24 nt rRNA fragments shows a noticeable gradual increase of the relative accumulation of 16-24 nt rRNA fragments in mock and TSA treatments (Figure 8.13). Conversely, the accumulation of 16 nt rRNA fragments was relatively suppressed in 4PBA treatment; the degree of suppression seems to display a negative association with the length of the fragment and the period of drug treatment. The difference between 4PBA and mock treatments in the accumulation level of 16-24 nt rRNA fragments seems to be inversely related with the size of rRNA fragment and period of 4PBA treated time. These data together indicate that mock treatment (which includes a would like treatment of the callus) stimulates rRNA fragment accumulation. The exposure of the callus to 4PBA appears to negate this stimulation in comparison to TSA, which appears to have little, if any, impact on rRNA fragment accumulation.

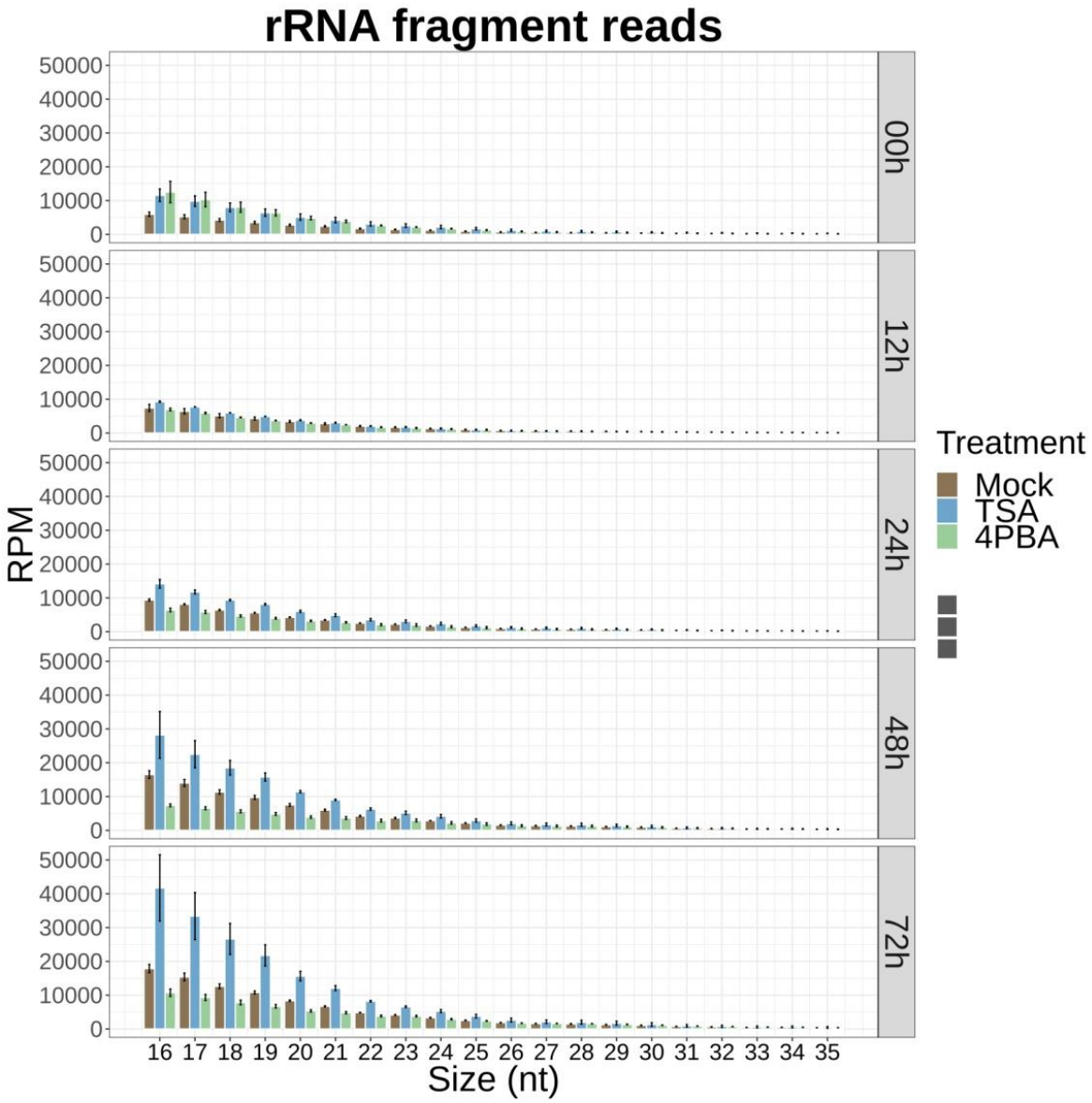



**Figure 8.12 The amount of rRNA fragments grouped by size**

rRNA fragments were grouped by size (x-axis), and the accumulation was shown as RPM (y-axis). The mean of three replicates ± SD was plotted. The mock, TSA and 4PBA treatments are shown in brown, blue and green colours, respectively. Time-point was as indicated on the right side of each panel.

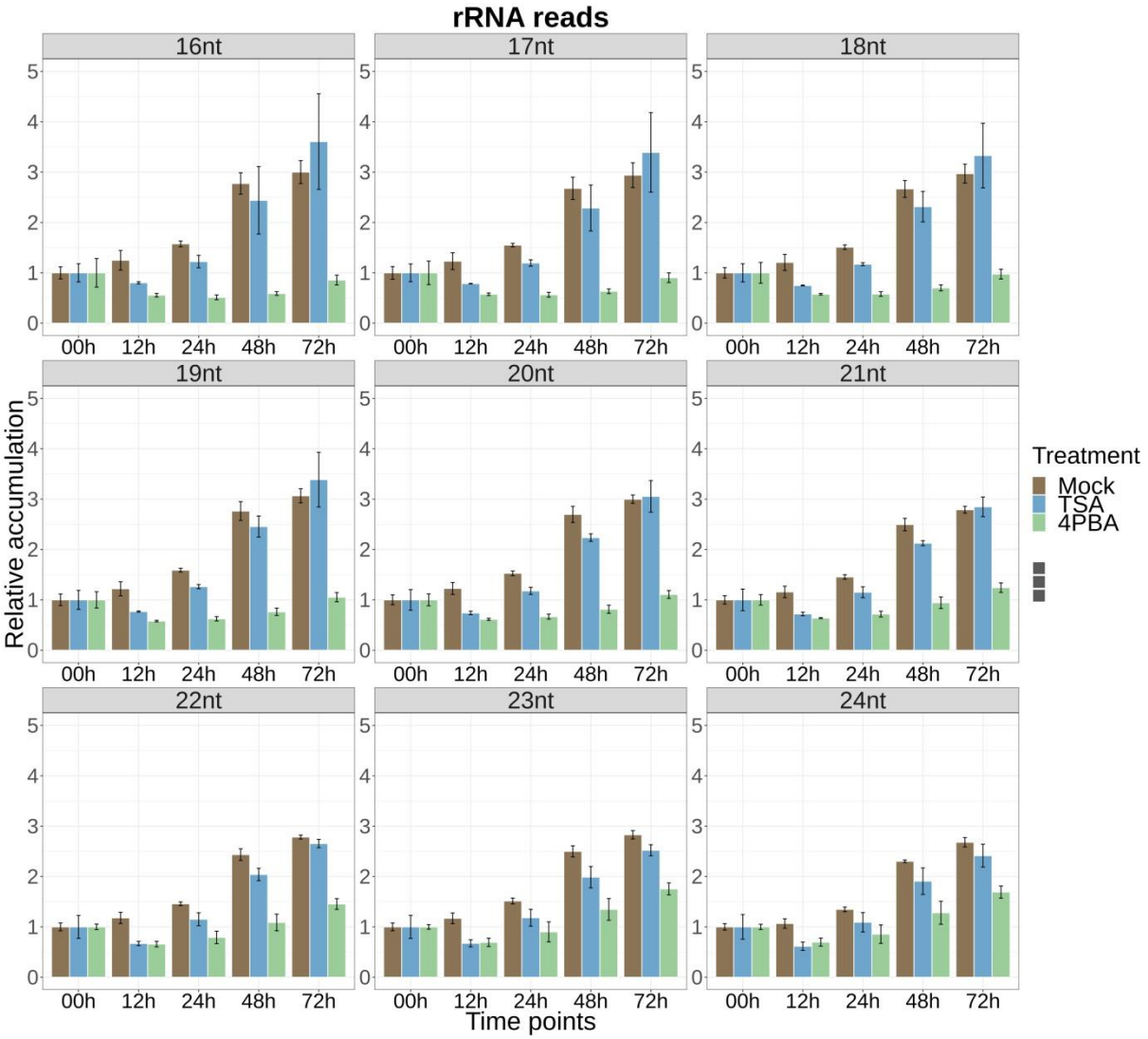



**Figure 8.13 Relative accumulation of 16-24 nt rRNA fragments**

The rRNA fragment RPM level of each size category was compared with the corresponding level at T=0. The relative accumulation (y-axis) was plotted against the time-point (x-axis). All data were shown as the mean of the three replicates ± SD.

### 8.4.2.3 miRNA

The miRNA level in mock treatment was generally as same as at T=0, whereas the miRNA level was slightly increased in 4PBA and slightly decreased in TSA (Figure 8.14). Although the differences in miRNA accumulation among treatment over time were statistically significant, these fluctuations were lower than half fold change, suggesting that the miRNA population was not heavily impacted. These miRNAs are mostly 20-22 nt in length (Figure 8.15), of which 21 nt miRNA was the most dominant category. As the relative accumulation of 21 nt miRNAs was rarely affected by the mock and HDACi treatments (Figure 8.16 B), the 20 nt and 22 nt miRNAs were increased by half in both 48 hours and 72 hours of 4PBA treatment (Figure 8.16), implicating possible changes in expression activity of individual miRNA genes, particularly those producing 20 nt and 22 nt miRNAs.

The accumulation level of distinct grapevine miRNA sequences collected and identified by Lizamore and Winefield (2017) were individually analysed. Among the 148 distinct miRNA sequences, 62 of them were present (> 10 reads) in at least one time-point of the treatments. Differential accumulation analysis revealed two miRNAs (vvi-miR3630-3p and vvi-miR396b) were significantly up-regulated in response to the mock treatment, and the rest of the miRNA was not significantly affected (Appendix C.15). The comparison of mock versus TSA or 4PBA treatments showed that these two miRNA products were not differentially accumulated in the TSA and 4PBA treatments. In fact, no miRNA was differentially accumulated in the comparison between TSA and mock treatment. The comparison between 4PBA and mock treatments revealed that, of the 11 miRNA sequences that were significantly and differentially expressed across the time course of treatment, two were up-regulated while the rest nine miRNA were down-regulated in the presence of 4PBA (Appendix C.15). Estimation of targeted TEs with stringent setting (E≤2) shows that only two of the significantly affected miRNA reveal a high affinity with three TE families; one is vvi-miR396b which was found to be up-regulated in mock, and the other is vvi-miR396d which was found to be down-regulated in the 4PBA treatment. Interestingly, these two miRNAs belong to the same miRNA family, and both preferentially target VLINE2, VLINE5 and Mutavine-17. It appears that silencing of VLINE2, VLINE5 and Mutavine-17 mediated by vvi-miR396b was enhanced in response to mock treatment. With the application of 4PBA adding to the mock treatment, the vvi-miR396b level was further slightly increased yet didn't exhibit over two-fold changes comparing with mock (Figure 8.17 A). Instead, the expression level of vvi-miR396d that preferentially targets the same TE families was significantly down-regulated in the presence of 4PBA (Figure 8.17 B), suggesting a milder PTGS silencing posed on VLINE2, VLINE5 and Mutavine-17 due to the presence of 4PBA.

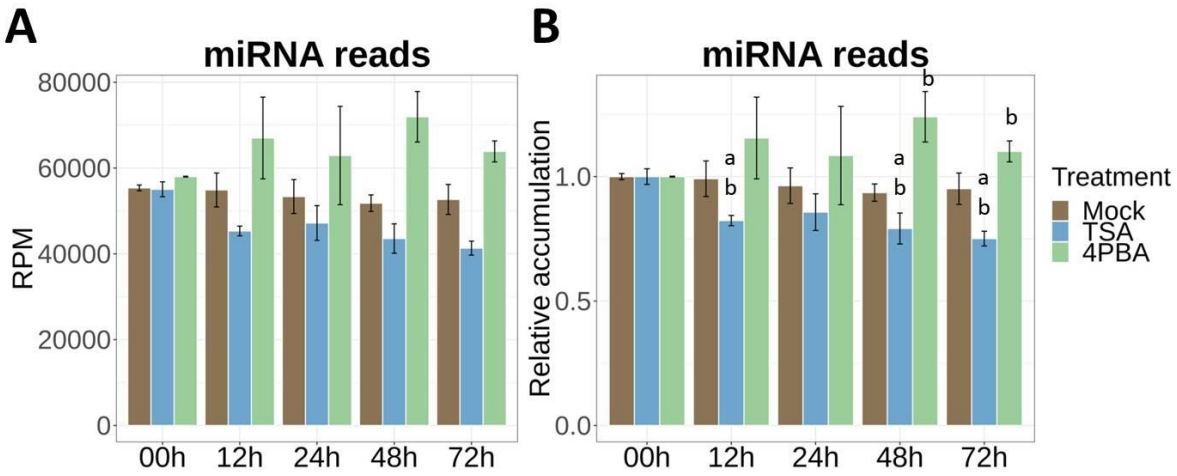



**Figure 8.14 The amount of miRNAs in multiple treatments over time**

**(A)** The accumulation of miRNAs shown as RPM. **(B)** The relative accumulation of miRNAs. All data were shown as the mean of the three replicates ± SD. The t-test was performed for the relative accumulation, where the comparisons versus T=0

(00h) with p-value < 0.05 are denoted by ‘a’ and the tests with p-value < 0.05 against mock at each time-point are denoted by ‘b’.

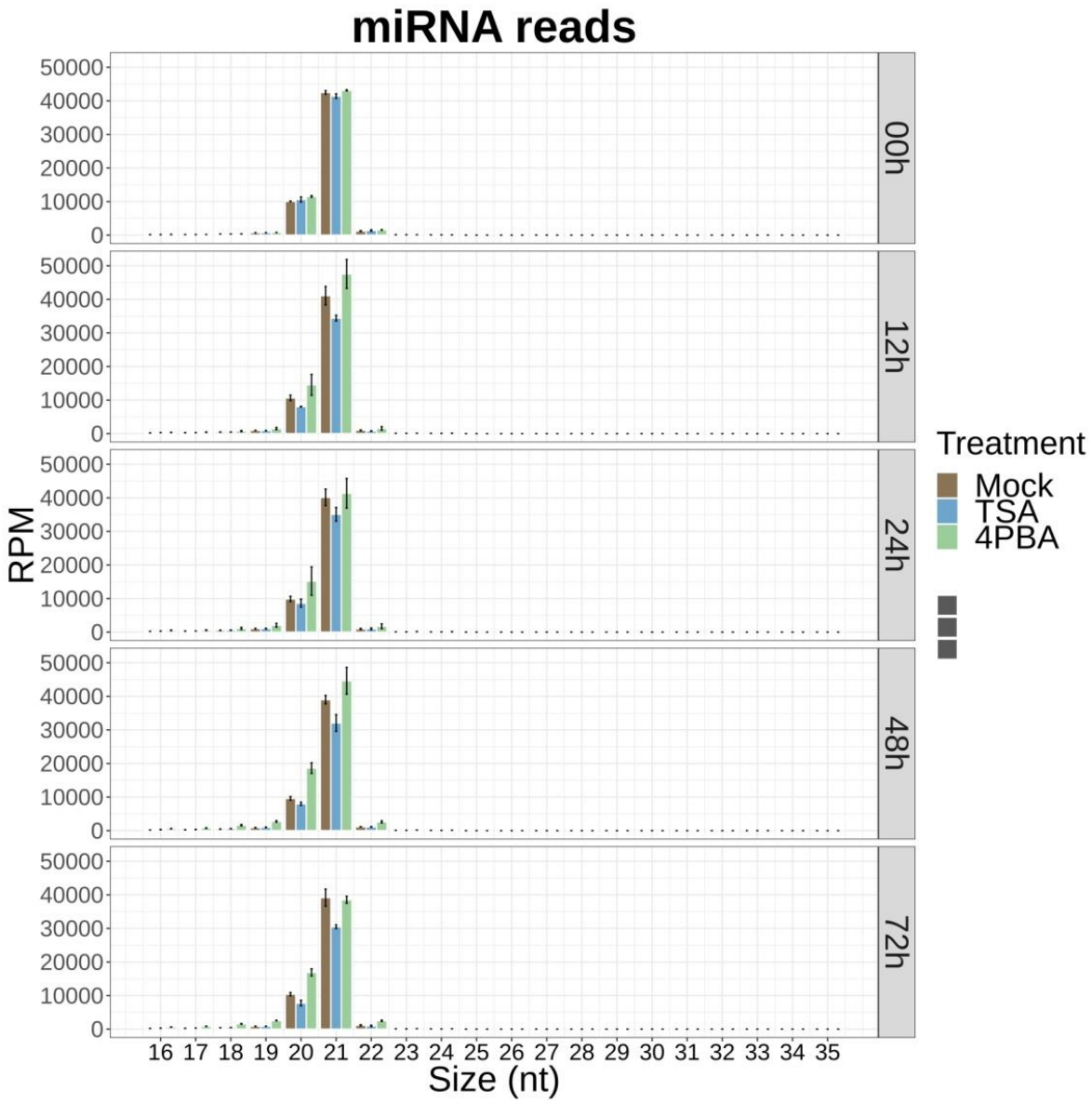



**Figure 8.15 The amount of miRNAs grouped by size**

miRNAs were grouped by size (x-axis), and the accumulation was shown as RPM (y-axis). The mean of three replicates ± SD was plotted. The mock, TSA and 4PBA treatments are shown in brown, blue and green colours, respectively. Time-point was as indicated on the right side of each panel.

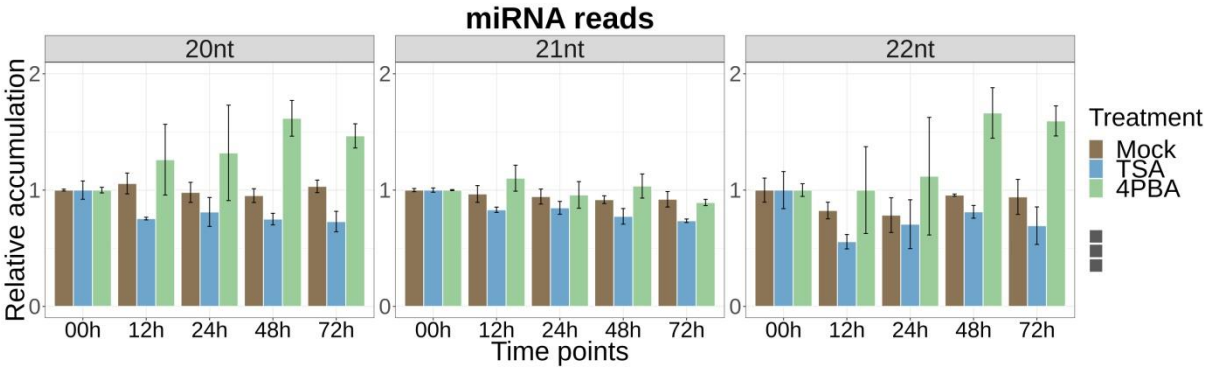



**Figure 8.16 Relative accumulation of 20-22 nt miRNAs**

The miRNA RPM level of each size category was compared with the corresponding level at T=0. The relative accumulation (y-axis) was plotted against the time-point (x-axis). All data were shown as the mean of the three replicates ± SD.

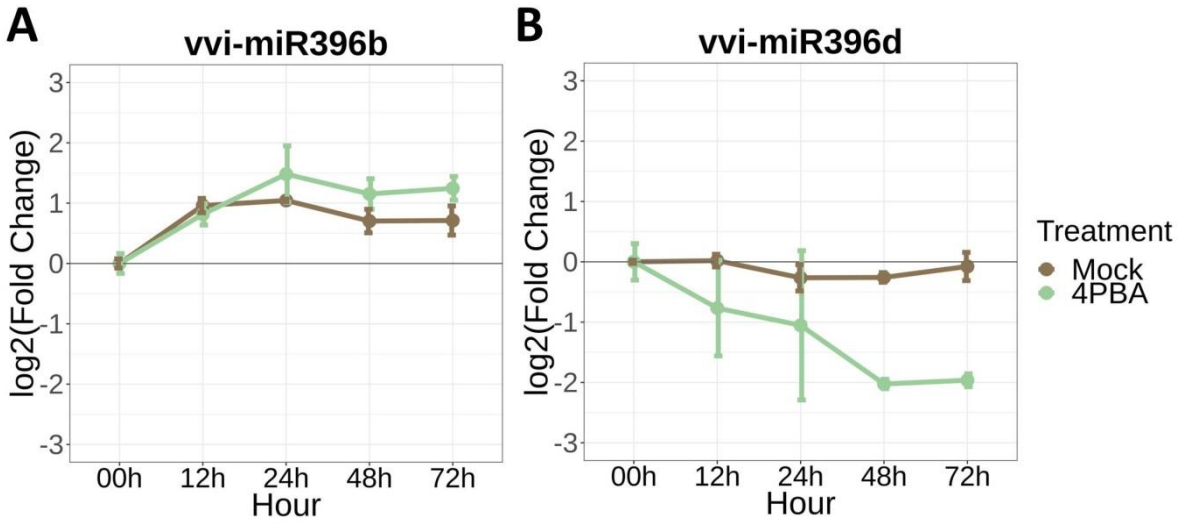



**Figure 8.17 Fold change of differentially expressed miRNAs**

The expression dynamics of **(A)** vvi-miR396b and **(B)** vvi-miR396d were presented in logarithmically transformed fold change (see 8.3 for details). All data were shown as the mean of the three replicates ± SD.

### 8.4.2.4 siRNA

Analysis of the relative accumulation of siRNA shows that the siRNA level was down-regulated in mock, and there was no apparent impact of the presence of 4PBA (Figure 8.18 A-B). The siRNA level was slightly suppressed in TSA treatment. Although these changes were statistically significant (p-value < 0.05), the dynamic changes of siRNA accumulation were minor (< 25%). Small interfering RNAs are generally 21-24 nt in length. In grapevine embryogenic callus, the majority of siRNA were 24 nt across all treatments (Figure 8.19), and siRNA accumulation was found to be relatively stable across all treatments and across all time points irrespective of the size category (Figure 8.20). Forty per cent to 50% of siRNA were derived from TEs across all treatments (Figure 8.21 A), and the

accumulation pattern of these siRNAs was similar to the overall pattern of the relative accumulation of all siRNAs (Figure 8.18 B, Figure 8.21 B). However, no significant differential accumulation of siRNAs derived from specific TE loci or TE families was observed, implicating trivial differences in the siRNA level across treatments over time.

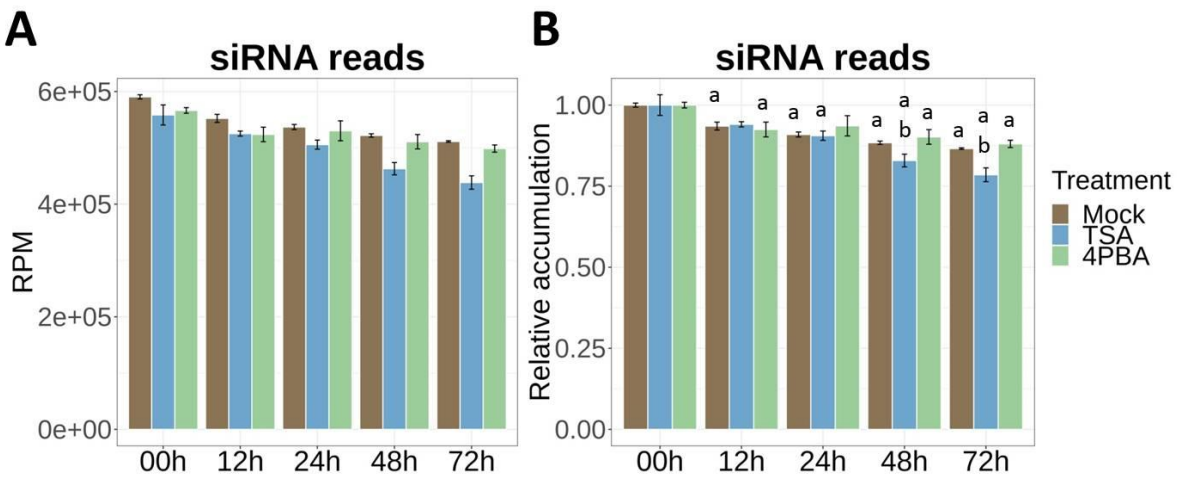



**Figure 8.18 The amount of siRNAs in multiple treatments over time**

**(A)** The accumulation of siRNAs shown as RPM. **(B)** The relative accumulation of siRNAs. All data were shown as the mean of the three replicates ± SD. The t-test was performed for the relative accumulation, where the comparisons versus T=0 (00h) with p-value < 0.05 are denoted by 'a' and the tests with p-value < 0.05 against mock at each time-point are denoted by 'b'.


**Figure 8.19 The amount of siRNAs grouped by size**

siRNAs were grouped by size (x-axis), and the accumulation was shown as RPM (y-axis). The mean of three replicates ± SD was plotted. The mock, TSA and 4PBA treatments are shown in brown, blue and green colours, respectively. Time-point was as indicated on the right side of each panel.

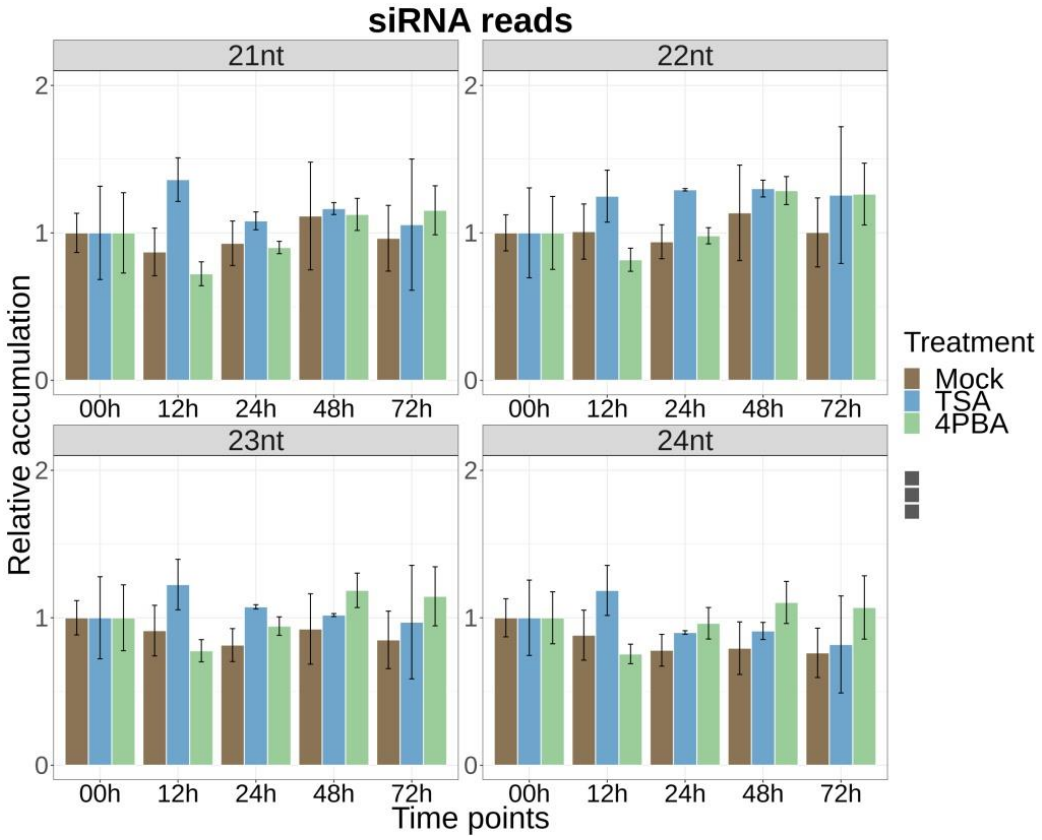



**Figure 8.20 Relative accumulation of 21-24 nt siRNAs**

The siRNA RPM level of each size category was compared with the corresponding level at T=0. The relative accumulation (y-axis) was plotted against the time-point (x-axis). All data were shown as the mean of the three replicates ± SD.

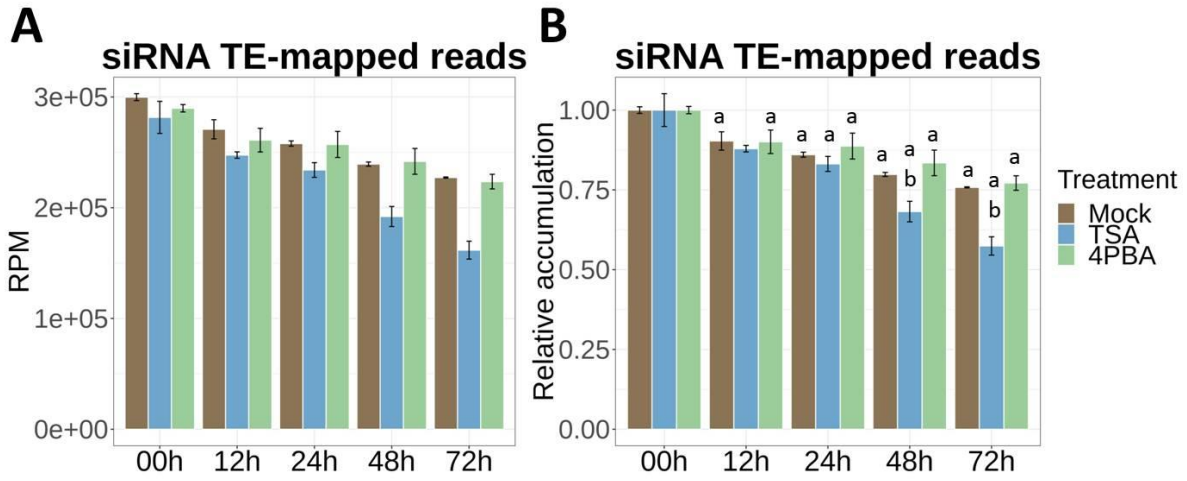



**Figure 8.21 The amount of TE-derived siRNAs in multiple treatments over time**

**(A)** The accumulation of TE-derived siRNAs shown as RPM. **(B)** The relative accumulation of TE-derived siRNAs. All data were shown as the mean of the three replicates ± SD. The t-test was performed for the relative accumulation, where the comparisons versus T=0 (00h) with p-value < 0.05 are denoted by 'a' and the tests with p-value < 0.05 against mock at each time-point are denoted by 'b'.

### 8.4.2.5 tasiRNA

A subset of siRNAs, named trans-acting small interfering RNAs (tasiRNAs), are processed in phase and acquire a trans-acting silencing behaviour. Trans-acting small interfering RNAs are generally 21-22 nt in length and indicative of amplified PTGS silencing signal. In the grapevine embryogenic callus, tasiRNA only contributed to about 1% of siRNA (Figure 8.22 A). The accumulation level of tasiRNA was relatively stable in mock and TSA treatments over time but was significantly down-regulated (~50%) in 4PBA treatment (Figure 8.22 B). The UEA Small RNA Workbench identified 280 tasiRNA loci producing 21-22 nt phased siRNA reads across all treatments over time. Intriguingly, only one locus (chr10_5280954_5281247) was differentially regulated and displayed three different response patterns in the three treatments. With a wide standard deviation, the expression level of this tasiRNA locus (chr10_5280954_5281247) in mock treatment was not significantly different from the level in T=0 (Figure 8.23 A). In the TSA treatment, this locus increased 16-fold compared to T=0. Conversely, 4PBA treatment led to the complete suppression (RPM=0) of this tasiRNA locus. This locus overlaps with an annotated TE, MULE-Mutavine-18_chr10_5277279-5286614. It is possible that the intermediate steps of MULE-Mutavine-18 tasiRNA biogenesis were affected in the presence of 4PBA, or it simply reflects reduced transcriptional activity of this TE family. Tracing back to the RNAseq data in chapter 7, the expression dynamics of the MULE-Mutavine-18 TE family shows that the overall transcriptional activity of this TE family was significantly down-regulated in 4PBA treatment, but not in mock and TSA treatments (Figure 8.23 B). These suggest that the significant reduction of the tasiRNA derived from MULE-Mutavine-18 is likely due to the decrease of the TE activity in 4PBA treatment.

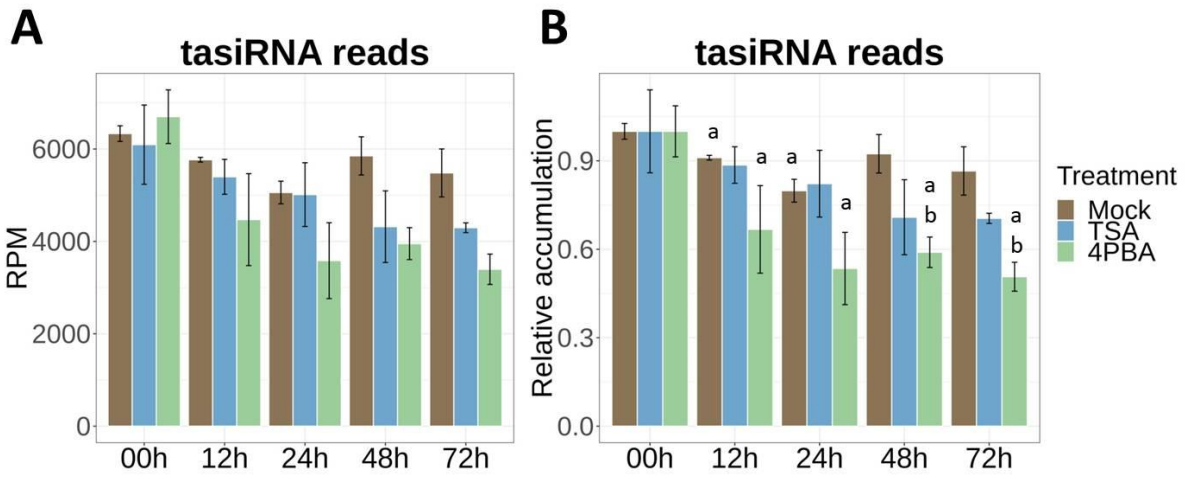



**Figure 8.22 The amount of tasiRNAs in multiple treatments over time**

**(A)** The accumulation of tasiRNAs shown as RPM. **(B)** The relative accumulation of tasiRNAs. All data were shown as the mean of the three replicates ± SD. The t-test was performed for the relative accumulation, where the comparisons versus T=0 (00h) with p-value < 0.05 are denoted by 'a' and the tests with p-value < 0.05 against mock at each time-point are denoted by 'b'.

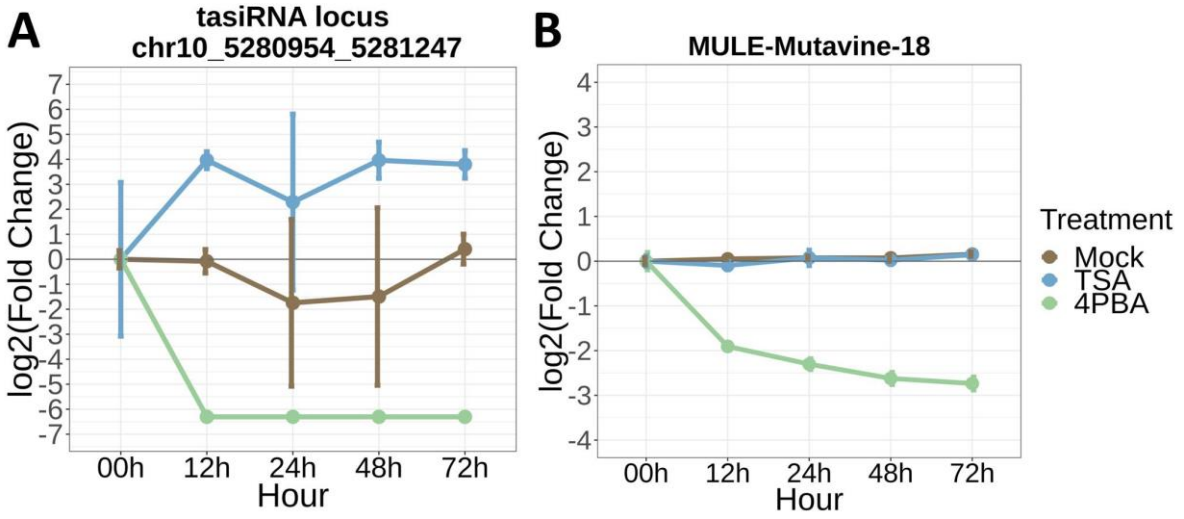



**Figure 8.23 Fold change of tasiRNA locus chr10_5280954_5281247 and MULE-Mutavine-18**

The expression dynamics of **(A)** tasiRNA locus chr10_5280954_5281247 and **(B)** MULE-Mutavine-18 family were presented in logarithmically transformed fold change. Data were shown as the mean of the three replicates ± SD.

### 8.4.3 Exploration of miRNA, siRNA and tasiRNA targeting or derived from Copia-3 and Copia-23.

From the findings in chapter 3, Copia-3 and Copia-23 are the two LTR-TE families that retain highly conserved autonomous loci and exhibit the most recent transposition burst in the grapevine genome. The Illumina RNAseq data revealed that some Copia-3 and Copia-23 loci were potential sites of autonomous transposition. However, interrogation of these samples using ONT full-length cDNA sequencing didn't detect contiguous full-length transcription from these autonomous loci (see chapter 6 and chapter 7). These suggest the level of autonomous TE transcripts was too low to be detected, or the epigenetic system still competently regulated TE activity, even with sporadic TE activation in response to stress or HDACi treatments. Except for tRNA and rRNA fragments, most of the sRNA categories, like miRNA, siRNA and tasiRNA, were maintained at a relatively stable level with at most half-fold changes compared with T=0. However, only a few differentially expressed miRNA and tasiRNA loci displayed a high affinity with TEs and appeared not to be differentially expressed. These together give rise to the assumption that most of the miRNA, siRNA and tasiRNA targeting or derived from TEs were maintained at a sufficiently abundant level to regulate TE mobility. To test this assumption, TE-related targets and origins of miRNA, siRNA and tasiRNA with over 100 RPM were investigated.

For miRNA, 29 distinct miRNA sequences were accumulated more than 100 RPM; five of them were predicted with strong specificity (E≤2, the smaller, the better) targeting TEs (Table 8.1). vvi-miR159c is the miRNA with the highest expression level conserved across all treatments over time. It preferentially targets Copia-23 and Copia-94, respectively, at 553 bp and 621 bp downstream of the PBS. Using NCBI's ORF finder to annotate the open-reading frame (ORF) of Copia-23 and Copia-94, the targeting sites are close to the 3' end of the first ORF encoding a protein of unknown function, whereas the predicted ORF2 encodes the poly-proteins (representative illustration for Copia-23 was shown in Figure 8.24 A). Although elevated transcriptional activity responding to 4PBA was observed in some individual autonomous Copia-23 loci, such as Copia-23_chr18_3274356-3279386 (Figure 8.24 B, C), the overall expression level of the Copia-23 family was two-fold decreased in 4PBA treatment compared with mock treatment (Figure 8.24 D). On the other hand, the Copia-94 expression level was stable in all treatments, and no differential expression was detected. Intriguingly, the grapevine gene VIT_211s0016g05010, which presumably encodes a detoxification gene (Glyoxalase I 7, GLYI7) responding to salinity stress (Pinedo et al., 2015), was also predicted as a target of vvi-miR159c with stringent complementary affinity (E=0.5). This gene was not differentially changed in mock and TSA treatments but significantly 16- to 32-fold up-regulated in 4PBA treatment (Figure 8.24 E). While vvi-miR159c was maintained at a high level steadily and estimated to target both VIT_211s0016g05010 and Copia-23 with high affinity, the up-regulation of VIT_211s0016g05010 and down-regulation of targeted Copia-23 seems controversial, suggesting that other factors mediating the PTGS pathway were also affected and that the gene locus VIT_211s0016g05010 and Copia-23 might be preferentially confronted by different epigenetic silencing pathways (see section 8.5.3).

**Table 8.1 Highly expressed miRNA (> 100 RPM) targeting TEs**

| miRNA ID | TE target (E≤2) | Average miRNA RPM from all libraries |
|---|---|---|
| vvi-miR159c | vvi-miR159c 21 AUCUCGAGGGAAGUUAGGUUU 1<br>:::.:::::.::::.:::::::<br>Copia-23 716 UAGGGCUCUCUUCGAUCCAAA 736<br>vvi-miR159c 21 AUCUCGAGGGAAGUUAGGUUU 1<br>:::.:::::.::::.:::::::<br>Copia-94 785 UAGGGCUCUCUUCGAUCCAAA 805 | 26923 |
| new_miR_03 | new_miR_03 21 UUCAAGUUCUUUCGACACCUU 1<br>.::::::: .::::::::::::<br>VLINE2 1638 UGGUUCAAAGAAGCUGUGGAA 1658<br>new_miR_03 21 UUCAAGUUCUUUCGACACCUU 1<br>:::: ::::::.:::::::::.<br>VLINE5 71 CAGUUGAAGAAGGCUGUGGAG 91<br>new_miR_03 21 UUCAAGUUCUUUCGACACCUU 1<br>::::::::::.:::::::: :<br>Mutavine-17 4661 GUGUUCAAGAAGGCUGUGGCA 4681 | 196 |
| vvi-miR396a | vvi-miR396a 21 AUCAAGUUCUUUCGACACCUU 1<br>:.::::::: .::::::::::::<br>VLINE2 1638 UGGUUCAAAGAAGCUGUGGAA 1658<br>vvi-miR396a 21 AUCAAGUUCUUUCGACACCUU 1<br>:::: ::::::.:::::::::.<br>VLINE5 71 CAGUUGAAGAAGGCUGUGGAG 91<br>vvi-miR396a 21 AUCAAGUUCUUUCGACACCUU 1<br>::::::::::.:::::::: :<br>Mutavine-17 4661 GUGUUCAAGAAGGCUGUGGCA 4681 | 150 |
| vvi-miR396b | vvi-miR396b 20 UCAAGUUCUUUCGACACCUU 1<br>:::: ::::::.:::::::::.<br>VLINE5 72 AGUUGAAGAAGGCUGUGGAG 91<br>vvi-miR396b 20 UCAAGUUCUUUCGACACCUU 1<br>.::::::: .::::::::::::<br>VLINE2 1639 GGUUCAAAGAAGCUGUGGAA 1658<br>vvi-miR396b 20 UCAAGUUCUUUCGACACCUU 1<br>::::::::::.:::::::: :<br>Mutavine-17 4662 UGUUCAAGAAGGCUGUGGCA 4681 | 137 |
| vvi-miR396d | vvi-miR396d 21 GUCAAGUUCUUUCGACACCUU 1<br>::::: ::::::.:::::::::.<br>VLINE5 71 CAGUUGAAGAAGGCUGUGGAG 91<br>vvi-miR396d 21 GUCAAGUUCUUUCGACACCUU 1<br>..::::::: .::::::::::::<br>VLINE2 1638 UGGUUCAAAGAAGCUGUGGAA 1658<br>vvi-miR396d 21 GUCAAGUUCUUUCGACACCUU 1<br>::::::::::.:::::::: :<br>Mutavine-17 4661 GUGUUCAAGAAGGCUGUGGCA 4681 | 10297 |

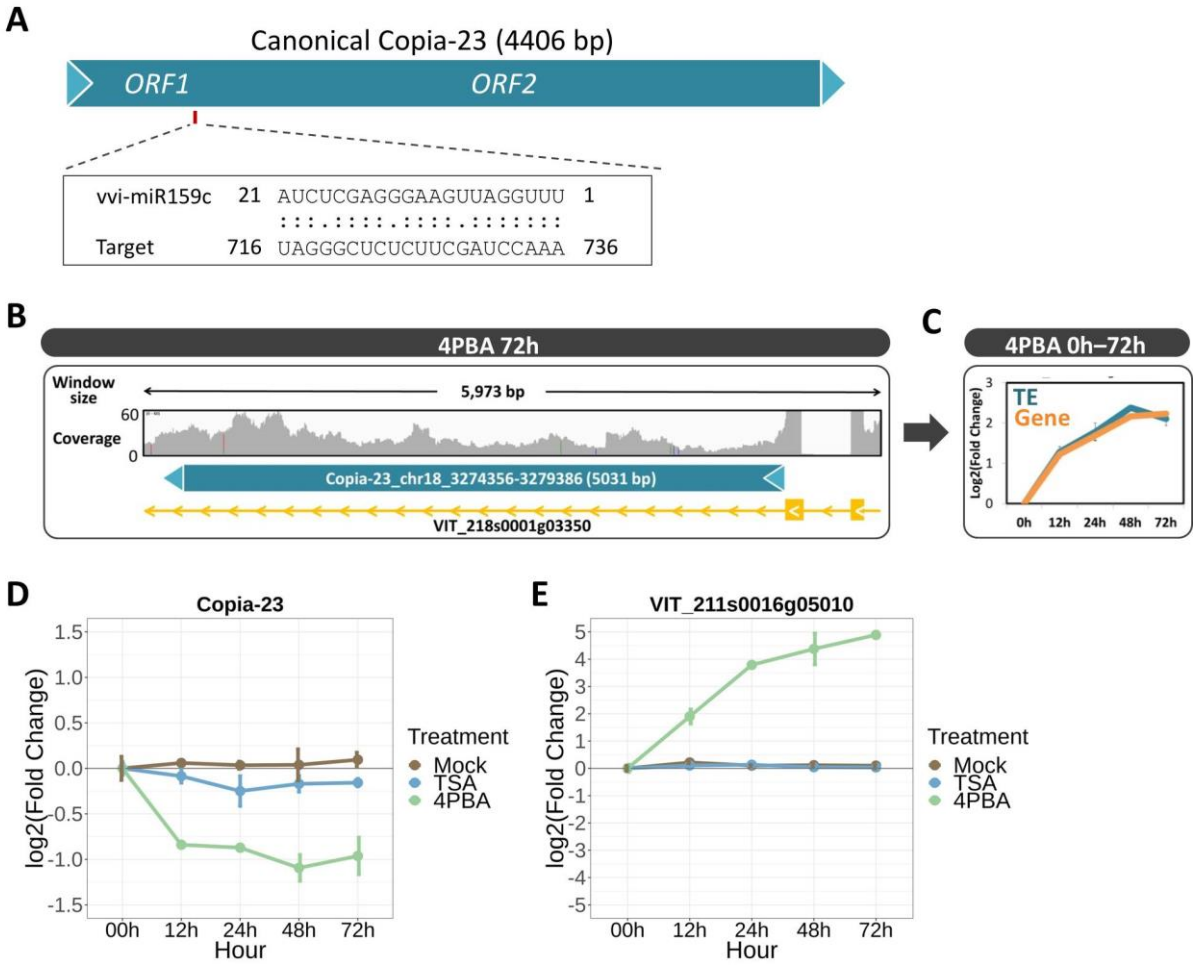



**Figure 8.24 TE and gene targeted by vvi-miR159c**

**(A)** Illustrated target site of vvi-miR159c on canonical Copia-23. **(B)** RNAseq read coverage of an up-regulated trackable Copia-23 locus. This TE locates within an intron of the indicated gene. **(C)** Logarithmically transformed expression dynamics of the TE and gene shown in (B). **(D)** The overall expression level of the Copia-23 family. **(E)** The expression level of VIT_211s0016g05010, which was also predicted to be targeted by vvi-miR159c. Note that this gene is not the gene locus presented in (B).

As introduced in 8.2, 24 nt siRNA are mostly through a Pol IV-DCL3-dependent pathway and involved in the maintenance of RdDM. The accumulation of primary and secondary 21-22 nt siRNA is usually indicative of TE activation; the former reflects direct degradation of TE mRNA, whereas the latter indicates signal amplification of PTGS and initiates *de novo* RdDM (Borges and Martienssen, 2015; Borges et al., 2018; Cuerda-Gil and Slotkin, 2016; Marí-Ordóñez et al., 2013). In grapevine embryogenic callus, only ~1% of total siRNA were 21-22nt secondary tasiRNA, while the majority of siRNA were 24 nt in length (Figure 8.19). Of the tasiRNA loci expressed at over 100 RPM, only one locus overlaps with an annotated TE MULE-MUDRAVI1. None of the expressed tasiRNA loci was found to overlap with annotated Copia-3 or Copia-23 loci, suggesting that no tasiRNA is likely to be generated from Copia-3 or Copia-23 in our system. Besides, as shown previously, the significantly down-regulated tasiRNA locus (chr10_5280954_5281247) in 4PBA treatment was derived from

MULE-Mutavine-18. Underrepresentation of 21-22 nt siRNA and lack of highly accumulated tasiRNA derived from Copia-3 and Copia-23 suggest that the PTGS was not considerably stimulated, possibly due to the non-threatening level of Copia-3 and Copia-23 transcripts. Analysing TE-derived siRNA by TE families revealed that 98 TE families produced siRNA at a level higher than 100RPM in at least one time-point (Appendix C.16). Among them, the average production of Copia-3-derived siRNA was ranked 10$^{th,}$ and Copia-23-derived siRNA was ranked 27$^{th}$, indicating that siRNA generated from these two families were maintained at an abundant level. As the majority of these siRNAs were 24 nt in length, this suggests that a reservoir of 24 nt siRNA was available for maintaining RdDM, negating the need for a significant new synthesis of 21-22 nt siRNA. In addition to this, grapevine genes potentially involving in RdDM maintenance were up-regulated in 4PBA treatment (see chapter 7). These genes are predicted to function as AGO2, RDR1 and NERD that mediate the Pol IV-NERD RdDM, as well as AGO4 that solely carry 24 nt siRNA to Pol V-transcribed transcripts and secure DNA methylation degree of the underlying sequences.

## 8.5 Discussion

### 8.5.1 Characteristics and potential roles of tRF in response to the wound-like treatment in grapevine

In addition to the well-known biological roles of tRNA in translation, researchers have seen the stress-induced accumulation of size-specific tRF in animals, plants and yeast, and this is not necessarily associated with impaired tRNA biogenesis (Thompson et al., 2008).

Various enrichment of tRFs in different lengths have been observed in *Arabidopsis* and rice. Generally speaking, *Arabidopsis* tRFs are predominantly 19-20nt in length, while rice tRFs are more represented by 25 nt fragments (Alves et al., 2017). In the grapevine embryogenic callus, tRFs are mostly represented by 16 nt fragments. In the grape system, we also observed the accumulation of 17-19 nt and 28-35 nt tRFs (Figure 8.5). Taken collectively, the observation of such a wide and variable distribution of fragment sizes may reflect a species-specific difference of tRF biogenesis (Alves et al., 2017).

It has been shown that mouse 3' CCA tRFs, particularly 18 nt and 22 nt in length, can widely target the conserved PBS site in LTR-TEs, thus causing inhibition of TE reverse transcription and translation via a miRNA-like mechanism (Schorn et al., 2017). It is known that transcriptional activation of endogenous retroviruses (ERVs) is accompanied by 3' CCA tRF accumulation in mouse embryonic stem cells (Schorn et al., 2017). However, in grapevine embryogenic callus with mock and HDACi treatments, 3' CCA tRF only comprised less than 1% of total tRFs measured (Figure 8.4 A, Figure 8.6 A), and no size category particularly outnumbered others (Figure 8.7) or significantly changed in the accumulated levels (Figure 8.8). Assuming 3' CCA tRNA has similar roles in plants as in mouse, the steady level of grapevine 3' CCA tRNA in all treatments suggests that the level of autonomous LTR-TE transcripts was not high enough to trigger extensive 3' CCA tRF biogenesis.

As described in section 8.2.3, some LTR-associated genes in mouse and Gypsy elements in *Arabidopsis* were found down-regulated or targeted by particular types of 5' tRF (Martinez et al., 2017; Sharma et al., 2016). In addition to TE suppression, by interacting with AGO or PIWI proteins, it is likely that 3' non-CCA tRF or 5' tRF can regulate gene expression through its miRNA- or piRNA-like properties reported in human, mouse, silkworm, rice and *Arabidopsis* (Alves et al., 2017; Honda et al., 2017; Keam et al., 2014; Kumar et al., 2014; Martinez et al., 2017; Schorn and Martienssen, 2018; Schorn et al., 2017). In the grapevine, the dynamic pattern of 3' non-CCA tRF was characterized with a peak around 12 hours or 24 hours post mock treatment and then gradually decreasing in the latter part of the time course (Figure 8.6 D). The additional continuous incubation with HDACi didn't alter this pattern of accumulation. These suggest that the response of 3' non-CCA tRF mainly reflected the

perturbation from the one-off wound-like treatment associated with the mock treatment, in which a wave of gene and TE transcriptional fluctuation has been observed in the Illumina RNAseq data (chapter 3, chapter 7). Therefore, an increase of grapevine 3' non-CCA tRF in mock treatment might be an epigenetic silencing response to regulate genes or TE expression. However, it requires a more thorough investigation to define particular types of 3' non-CCA tRF and targeted TEs or genes.

### 8.5.2 Accumulation of smallerrRNA fragments resembles non-specific degradation

Size-specific degradation of rRNA has been reported in stress-treated yeast and oat (Hoat et al., 2006; Mroczek and Kufel, 2008). However, according to these studies, rRNA fragments are more likely to be related to cell death signalling rather than epigenetic silencing. Depending on the types of stress sources, toxins or chemicals, various degradation patterns of rRNA can be generated. These degradation patterns are a combination of rRNA fragments sizing from 300 nt to 600 nt. In oat leaf, 200 nt-specific cytosol rRNA fragmentation was linked with programmed apoptotic cell death induced by the fungus toxin victorin or sodium azide ($NaN_2$), whereas random-size degradation with smeared pattern accumulated towards size below 200nt was linked to necrotic cell death triggered by copper sulphate ($CuSO_4$) or heat (Hoat et al., 2006). By examining rRNA fragments sizing between 16 nt to 35 nt, a significant gradual increase of rRNA fragments was observed in grapevine embryogenic callus treated with mock and TSA over time (Figure 8.11 B). The rRNA fragment accumulation seems positively correlated with the time period. Besides, the size distribution of these fragments resembles that commonly associated with degraded total RNA (Figure 8.12; Cholet et al., 2019). Intriguingly, this phenomenon was significantly improved or even prevented by the application of 4PBA. This might associate with the cytoprotective properties of 4PBA previously described in chapter 7.

### 8.5.3 Loosened PTGS but strengthened RdDM

In the *Arabidopsis* wild-type model, the transcriptional activity of TEs is significantly increased in pollen compared with inflorescence (Borges et al., 2018). This TE activation is accompanied by significant up-regulation of 21-22 nt siRNAs derived from the transcriptionally activated TEs (Borges et al., 2018). However, in our study, TE-related 21-22 nt siRNAs (tasiRNAs) were rare, and most of the differentially accumulated miRNA did not seem to target TEs preferentially, suggesting that the 4PBA-induced activation of TE transcription observed in chapter 7 was not high enough to stimulate significant up-regulation of 21-22 nt siRNAs or miRNAs. Although the accumulation level of miRNAs and siRNAs was relatively stable in grapevine embryogenic callus irrespective of treatments, the efficiency of PTGS and RdDM is likely to be affected by differentially expressed genes encoding proteins involving in these pathways (discussed as follows).

It has been reported that DCL2 and DCL4 participate in the production of 21-22nt primary siRNAs, while AGO1 interacts with 21-22nt primary siRNAs and miRNAs and direct these sRNAs to the complementary mRNA, triggering cleavage at the target site (Borges and Martienssen, 2015; Cuerda-Gil and Slotkin, 2016). In our study, although the miRNA and 21-22 nt siRNA (including tasiRNA) were generally maintained at a steady level in all treatments, a reduction in AGO1 and DCL2 expression observed in 4PBA treatment (Figure 7.15) may negatively affect the efficiency of siRNA and miRNA deployment, thus weaken PTGS.

In contrast, AGO2, AGO4, RDR1, RDR6 and NERD, key proteins to canonical and non-canonical RdDM pathways (Figure 8.1; section 8.2.4), were found to be up-regulated in 4PBA treatment in our system (Figure 7.15). Providing an abundant and stable level of 24 nt siRNA in the embryogenic callus subjected to mock and 4PBA treatments (Figure 8.19 and Figure 8.20), increased concentration of AGO2, AGO4, RDR1, RDR6, and NERD in 4PBA treatment might re-enforce suppressive DNA methylation and histone modification on the existing heterochromatic gene or TE loci.

Besides, several TE-derived siRNAs, mostly 24 nt in length, accumulated at a high and stable level with > 100 RPM (Appendix C.16). The TE families associated with these 24 nt siRNAs include Copia-3 (ranked 10) and Copia-23 (ranked 27), the two LTR-TE families considered to be most competent for autonomous transposition (see chapter 3, chapter 7). In our study, since numerous TE-derived 24 nt siRNA was maintained at a high level throughout the time period of the treatments, and the up-regulation of aforementioned genes involving canonical/non-canonical RdDM pathways might lead to more efficient utilization of the 24 nt siRNA reservoir, it is likely that the 24 nt siRNA-driven RdDM is the key goalkeeper for the suppression of activated TE transcription in grapevine callus subjected to HDAC inhibition by 4PBA treatment.

Instead, the most highly and stably expressed miRNA, vvi-miR159c (Table 8.1), was predicted to preferentially target Copia-23 and Copia-94, two TE families expressed at high (1100-1200 FPKM) and medium (500-600 FPKM) level in the treated callus, respectively. However, in the presence of 4PBA, the silencing effect of vvi-miR159c was likely to be compromised by down-regulated AGO1 in our system since it has been reported that miR159 can be loaded onto AGO1 in *Arabidopsis* (Qi et al., 2005).

A surprising finding here is the discrepancy in the expression changes of a gene (VIT_211s0016g05010) and TE (Copia-23) that were estimated to be both targeted by vvi-miR159c. The gene VIT_211s0016g05010 was predicted to be a hotspot of vvi-miR159c targeting and was 16-

to 32-fold up-regulated in 4PBA treatment, whereas the overall expression level of Copia-23 was suppressed (Figure 8.24). It seems that, with stably expressed vvi-miR159c, the epigenetic suppression on VIT_211s0016g05010 might be weakened in 4PBA treatment as the carrier of vvi-miR159c, i.e. AGO1, was down-regulated. If this were to be the case, this could have risked the host cell with elevated Copia-23 transcripts. However, it appears that the host cells might be able to prevent this from happening through re-enhancing RdDM on Copia-23 loci with up-regulated AGO2, AGO4, RDR1, RDR6 and NERD (Figure 8.25).

Taken together, based on the results in chapter 7 and chapter 8, an adjusted model of the homeostasis between sRNA and TE activity is proposed (Figure 8.26). In the wild-type background, the materials for RdDM, 24 nt siRNAs, are maintained at a stable but ample level. When there is a TE storm initiated by stress or chemical perturbation, the factors responsible for the deployment of 24 nt siRNA would be up-regulated to strengthen the transcriptional silencing circuit, i.e. RdDM. As a consequence, the overall transcriptional activity of TEs would be suppressed, albeit some sporadic activation at individual TE loci co-localized with active genes. Since the overall TE activity is down-regulated, the host cells might be able to de-escalate the PTGS system, which in turn can benefit genes co-regulated by PTGS and facilitate stress response. Another characteristic of this proposed model is that miRNAs and siRNAs are maintained at a relatively stable and plentiful level in the wild-type background to make sure the epigenetic system is unlikely to be compromised in the majority of cases. Instead, adjusting the activity of genes and proteins participating in specific epigenetic pathway might provide more flexibility and variability in response to various conditions or threats.

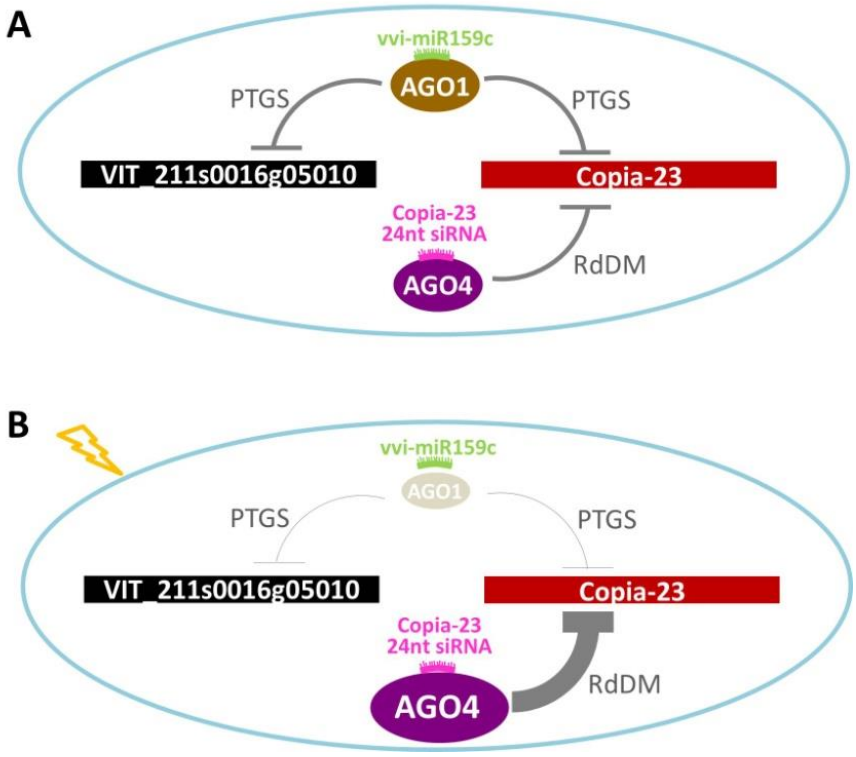



**Figure 8.25 Proposed model for TE and gene co-regulated by the same miRNA**

**(A)** In the wild-type background, grapevine gene VIT_ 211s0016g05010 is regulated by highly accumulated vvi-miR159c, which also post-transcriptionally silences Copia-23. In addition to the regulation from the PTGS pathway, Copia-23 is also monitored by RdDM mediated by 24 nt siRNAs and AGO4. **(B)** With the presence of 4PBA, a reduced level of AGO1 resulted in loosened PTGS, thus allow the accumulation of VIT_ 211s0016g05010 for the stress response. Given the excessive level of vvi-miR159c and non-depleted level AGO1, the weakened PTGS might still be able to sequester sporadic Copia-23 transcripts. In addition, an elevated level of AGO4 can direct Copia-23 siRNA to the TE loci and re-enhance RdDM. The down-regulation of TE activity due to strengthened RdDM might further signal to the epigenetic machinery for further de-escalation of PTGS.

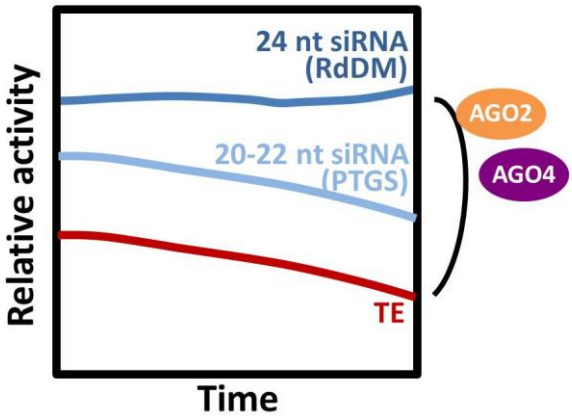



**Figure 8.26 Proposed model of the balance between sRNAs and TE activity**

In a wild-type background, 24 nt and 20-22 nt siRNAs are both maintained at an excessive level. In the presence of stress, the accumulation level of 24 nt siRNAs is not changed over time. Instead, transcriptional activation of AGO2 or AGO4 increases the use of 24 nt siRNA in targeting TEs and enhancing RdDM. This might, in turn, result in de-escalation of PTGS mediated by 20-22 nt siRNAs.

## 8.6 Conclusions

The findings in this chapter show that the TE perturbation caused by the mock treatment, which provides a wound-like pre-treatment, and HDACi incubation did not overwhelm the epigenetic system. In fact, with a stable and ample level of 24 nt siRNAs, the RdDM system was likely to be enhanced by the up-regulated AGO2, AGO4, RDR1, RDR6, and NERD in 4PBA treatment. Providing that the overall transcriptional activity of TEs was inhibited by strengthened RdDM, it is possible that the PTGS system involving 20-22 nt miRNAs, as well as primary and secondary siRNAs, could be de-escalated by down-regulation of AGO1 rather than a dramatic decrease in sRNA level. Furthermore, the discrepancy in the transcriptional level of gene and TE both targeted by vvi-miR159c provides an insight into how the crosstalk between PTGS and RdDM can de-repress gene activity in response to stress but not expose the host cell to extensive TE re-activation.

# Chapter 9

# Overall Conclusion

## 9.1 Reviews of the hypotheses

### 9.1.1 $H_1$: It is possible to distinguish a subset of transcriptionally active TE loci.

In chapter 2, an analysis pipeline was built for collecting potentially expressed TE loci with the conventional RNAseq data. By harnessing multiple existing tools, this pipeline was able to narrow down the search area from 223,411 annotated TE loci to about 3,700 expression candidates in non-treated grapevine embryogenic callus and about 5,000-5,500 expression candidates in stress-treated callus. In chapter 3, these expression candidates were further sorted by the presence or absence of unique-mapping reads; TE loci fell into the former were trackable while the rest were untrackable. As trackable loci were indicative of increased divergence over evolutionary time, the untrackable loci represented highly conserved sequences, possibly implicating more recent transposon activity. Among the 232 TE families in grapevine, two LTR-TE families, Copia-3 and Copia-23, were over-represented in the category of full-length untrackable expression candidates, particularly in the collection of potential origins of autonomous transcripts, suggesting these two families are most likely seeding new transposition in this tissue and under the treatments explored.

Application of the analysis pipeline on the *Arabidopsis* and *Drosophila* RNAseq data (chapter 5) also successfully extracted a subset of potentially expressed TEs from the inactive loci, despite the resolution of the expression candidate pool in terms of trackable/untrackable proportion being apparently species-specific.

Taken together, this study has developed a valuable approach to identify a remarkably reduced set of transcriptionally active TE loci, and facilitate interrogating the expression of TEs within or near gene sequences, therefore providing an implication for the determination of the epigenetic role of TEs in modulating gene activity. Overall, the results in this study disprove the null hypothesis:

*$H_{0\text{-}1}$: It is not possible to tell the difference between transcriptionally active TE loci and the silenced ones.*

### 9.1.2 $H_2$: The position of TEs within genes can reveal the transcriptional activity of TEs

Of the grapevine embryogenic callus, the strong location bias toward introns of expressed genes gives rise to the suggestion that TE loci position in the intron of active or inducible genes are more

likely to have access to transcription machinery, possibly due to the relatively relaxed chromatin conformation for gene activation. In addition, the autonomous loci likely to produce full-length Copia-3 and Copia-23 transcripts were mostly within the intron of expressed genes.

In the later stages of this research, ONT cDNA sequencing was available for the validation of the observations from short-read sequencing data and investigation of full-length TE transcripts derived from de novo transcription of TEs or from the transcription of neighbouring gene loci. Several structurally autonomous TE loci of Copia-3, Copia-23, and LINE elements exhibiting over 90% breadth of coverage of ONT sequencing read. However, a thorough investigation revealed that most of the nearly full-length coverage was contributed by overlapped ONT reads, each of which could not account for an intact full-length transcription of the autonomous TE loci. This suggests that these ONT reads represented transcription that was started from or stopped at the cryptic transcription start or stop sites within TE loci, respectively, or these ONT reads might be actually derived from fragmented TE loci that were identical to part of the sequence of the structurally autonomous TE loci. Among all TE families in grapevine, only Gypsy-V1 and hAT-7 showed a low level of full-length transcripts in the embryogenic callus subjected to mock treatment, which comprises a procedure resembling wound treatment. Although the analysis in chapter 6 did not detect intact full-length transcription of Copia-3 and Copia-23, the results in chapter 6 validated the location bias of expressed TE loci toward intron of expressed genes (chapter 3).

Analysis of the *Arabidopsis* RNAseq data showed similar location bias of expression candidates in wild-type and *ibm2*. On the contrary, *ddm1* revealed a substantial increase in intergenic proportion compared with wild-type and *ibm2* (chapter 5). Taking together, TE within introns of expressed genes might take advantage of the permissive chromatin conformation granted for gene activation, which could have in turn resulted in insertion bias of Copia-3 and Copia-23. Unless the epigenetic system is compromised, intergenic TEs have less chance to be stimulated. Notably, this model is likely not the case in *Drosophila*.

Although this model is likely not the case in *Drosophila*, the findings in grapevine and *Arabidopsis* disprove the null hypothesis:

*$H_{2\text{-}0}$: The position of TEs within genes is not associated with the transcriptional activity of TEs.*

### 9.1.3 $H_3$: The transcriptional activity of intragenic TEs is associated with the activity of the corresponding host genes.

It has been reported that genes proximal to or containing TE insertions tend to be transcribed at lower expression level than are genes distal to TEs or without intragenic TE insertions (Hollister and Gaut, 2009; Le et al., 2015), suggesting that presence of TEs within or near genes has a negative

effect on the expression of co-localized genes. However, there are also studies that found co-activation of TEs and co-localized genes, in terms of transcriptional level, in *Arabidopsis* subjected to biotic stress or nutrient starvation (Dowen et al., 2012; Secco et al., 2015). In our system, we interrogated both phenomena that seem controversial with each other.

Using both Illumina Truseq RNAseq and ONT cDNA sequencing data, genes co-localizing with expression candidates, particularly full-length TEs, were found less likely to be highly expressed than genes without TEs. On the other hand, if all the intragenic TE insertions were not expressed, the host genes were not likely to be highly expressed either. These imply a suppressive effect preferentially on genes co-localized with TEs, whereas genes need to be maintained with high transcripts level are less likely to co-localize with TEs. Nevertheless, the expression dynamics of differentially expressed TEs (DETEs) were mostly concordant with that of co-localized differentially expressed genes (DEGs). We hypothesized that the 'trade-off' theory proposed by Hollister and Gaut (2009) could link the negative effect of the presence of co-localized TEs on gene expression activity and the concordant expression pattern of co-localized DETEs and DEGs. It is possible that genes co-localized with TEs were prohibited from high transcriptional activity to prevent high expression of co-localized TEs. However, because the expression of some genes co-localized with TEs was required for the host cells, the chromatin structure may not be tightly packed, and the transcription machinery was likely not depleted around these regions and thus permitting and tolerating the transcription of hitchhiking TEs. Therefore, since the DEGs and co-localized DETEs may be under the same chromatin status, their expression patterns would be consistent.

Overall, these observations disprove the null hypothesis:

*$H_{3\text{-}0}$: No significant relationship exists between the transcriptional activity of TEs and corresponding host genes.*

### 9.1.4 $H_4$: Inhibition of HDACs that are key enzyme to maintain compact chromatin structure can facilitate TE re-activation.

Although the total numbers of expression candidates were not increased in the presence of HDACi, TSA and 4PBA, there was a wider range of TE families contributing expression candidates with the presence of 4PBA. Besides, the proportions of intergenic expression candidates and those in the flanking regions were increased in the 4PBA treatment, suggesting a shift of permissive transcriptional areas. Besides, a substantial amount of TE loci were significantly up-regulated in 4PBA treatment but not in the mock or TSA treatments. Many of these stimulated loci were inactive in mock and TSA treatments but newly emerged in the presence of 4PBA. These findings not only show that 4PBA seems to be more effective than TSA but also demonstrate a 4PBA-induced TE perturbation that might give rise to autonomous TE mobilization. Nonetheless, using ONT cDNA

sequencing, we were unable to detect intact full-length transcripts derived from structurally autonomous TE loci. It is likely that HDACi, especially 4PBA, can result in a perturbation of TE transcription, in which considerable up-regulation of TE loci were observed in the Illumina Truseq RNAseq data. However, it might require both proper stimuli corresponding to the properties of CREs in TEs and a compromised epigenetic system to extensively trigger TE transcription detectable by ONT cDNA sequencing and also facilitate TE mobilization. From the perspective of transcriptional activity, but not autonomous full-length transcription, 4PBA did elevate the transcription level of a subset of TE loci derived from various TE families. Therefore the null hypothesis $H_{4\text{-}0}$ is rejected.

*$H_{0\text{-}4}$: TE transcriptional activity cannot be stimulated by inhibition of HDACs.*

### 9.1.5 $H_5$: TE perturbation due to HDACi can, in turn, enhance PTGS or RdDM.

In grapevine embryogenic callus with mock pre-treatment and incubation of 4PBA, the TE perturbation was accompanied with up-regulation of RdDM factors, e.g. AGO4, AGO2, RDR1, RDR6, and NERD, as well as down-regulation of PTGS key factors, e.g. AGO1 and DCL2. The accumulation level of miRNAs and TE-derived siRNAs, mainly 24 nt siRNAs, were maintained at an ample level. The two most recently active Copia-3 and Copia-23 produced an extensive amount of 24 nt siRNAs above 100 RPM. Combining the findings of the high level of TE-related 24 nt siRNA and the differential expression of the aforementioned PTGS and RdDM factors, it seems that the RdDM pathway was enhanced whereas PTGS was down-played. This might explain the lack of competent TE transcripts. In addition, the most highly expressed miRNA, vvi-miR159c, was presumed preferentially targeting both Copia-23 and a gene (VIT_211s0016g05010) predicted to encode a glyoxalase enzyme. Surprisingly, as the overall transcriptional level of Copia-23 was down-regulated, possibly due to enhanced RdDM, the expression level of the vvi-miR159c-targeted gene (VIT_211s0016g05010) was 16- to 32-fold elevated in 4PBA treatment. The up-regulation of VIT_211s0016g05010 may be related to the down-regulation of AGO1, which is presumably required for carrying vvi-miR159c to target the gene transcripts. These together give rise to the proposed model that, in the wild-type background, RdDM is the major epigenetic pathway to be escalated in responding to a transcriptional perturbation of endogenous TEs if the transcriptional level of TEs was not high enough to trigger an extensive accumulation of 21-22 siRNAs as a signal of PTGS. In turn, the PTGS pathway may be suppressed to allow transcriptional activation of stress-responsive genes that are regulated by miRNAs. In this model, miRNAs and siRNAs are maintained at a far more excessive level than needed. In this way, the host cells might be able to fine-tune epigenetic pathways flexibly by adjusting the activity of various epigenetic-related genes. These observations, therefore, disprove the following null hypothesis:

*$H_{0\text{-}5}$: HDACi-induced TE perturbation has no effect on PTGS and RdDM.*

## 9.2 Future work

Following the conclusions from chapter 8, it is likely that the DNA methylation level would be maintained or even increased at TE loci. With whole-genome bisulfite sequencing, there will be a chance to examine the proposed model of enhanced RdDM. In addition, it would require more experiments to test whether vvi-miR159c can indeed cause cleavage of Copia-23 and the targeted gene transcripts. Evidence of the interaction between vvi-miR159c and Argonaute proteins, e.g. AGO1 or AGO10, will further strengthen the hypothesis. Providing the proved co-regulation of Copia-23 and the Glyoxalase-encoding gene by vvi-miR159c, phylogenetic analysis of the miR159c homologs and their gene or TE targets across different plant species might shed light on the evolution of miRNAs regulating both genes and TEs.

While gene and TE transcriptional activity can be inferred from short-read sequencing, it is necessary to include long-read based methods to recapitulate the transcription information with high continuity. With improved chemistry and base-calling algorithms, the ONT platform can be widely applied for TE-oriented studies. Recently, Lee et al. (2020) harnessed ONT DNA sequencing to detect intact cDNA molecules enclosed by the virus-like particles (VLPs) of autonomous LTR-TEs. As these cDNAs are the intermediates immediately before the insertion into genomic sequences, their presence presents more conclusive evidence of LTR-TE mobilization than the presence of full-length transcripts. Providing the proper combination of stress treatment and chemical inhibition of key points in the epigenetic silencing pathways, e.g. histone deacetylation and DNA methylation, to stimulate TE transposition, the examination of TE activity should be conducted using the VLP- and ONT-based approach to acquire confident evidence of the elements likely to be able to transpose.

Although the establishment of tissue culture like embryogenic callus has been considered introducing new TE insertions (Lizamore, 2013; Rakocevic et al., 2009), the transposition efficiency seems low. In addition, while TE mobilization was observed in somatic tissues, few were transmitted to progeny unless the mutation took place in reproductive tissue (Ito et al., 2011). The naturally occurring chimeric grape berries of different colours and the colour pigmentation on the morning glory petals due to spontaneous TE mobilization suggest mutation took place in the reproductive meristems, possibly associated with the developmental relaxation of transposon silencing in plants (Inagaki et al., 1994; Martínez and Slotkin, 2012). Therefore, for the long-term goal of introducing genetic diversity to clonal crops, stimulation of TE mobilization in the meristematic tissues is likely to be more effective than using other types of somatic tissues. Furthermore, this might also provide a platform to question the manner in which TEs are amplified in natural systems.

To our knowledge, this research is the first to demonstrate the properties and transcriptional landscape of expressed TE loci in relation to genes at a genome-wide scale. Our findings suggest a positive cycle of TE proliferation strategy, in which TE insertions prefer the transcribed chromatin region, which, in turn, would facilitate TE transcription. Besides, this research has established several TE-oriented analysis workflows that are applicable in different species to interrogate TE biology and test our ideas.

# References


Alao, J.P., Stavropoulou, A.V., Lam, E.W.-F., and Coombes, R.C. (2006). Role of glycogen synthase kinase 3 beta (GSK3β) in mediating the cytotoxic effects of the histone deacetylase inhibitor trichostatin A (TSA) in MCF-7 breast cancer cells. Mol. Cancer *5*, 40.

Alinsug, M.V., Yu, C.-W., and Wu, K. (2009). Phylogenetic analysis, subcellular localization, and expression patterns of RPD3/HDA1 family histone deacetylases in plants. BMC Plant Biol. *9*, 37.

Allen, R.S., Li, J., Stahle, M.I., Dubroue, A., Gubler, F., and Millar, A.A. (2007). Genetic analysis reveals functional redundancy and the major target genes of the Arabidopsis miR159 family. Proc. Natl. Acad. Sci. *104*, 16371–16376.

Alló, M., Buggiano, V., Fededa, J.P., Petrillo, E., Schor, I., De La Mata, M., Agirre, E., Plass, M., Eyras, E., Elela, S.A., et al. (2009). Control of alternative splicing through siRNA-mediated transcriptional gene silencing. Nat. Struct. Mol. Biol. N. Y. *16*, 717–724.

Alves, C.S., Vicentini, R., Duarte, G.T., Pinoti, V.F., Vincentz, M., and Nogueira, F.T.S. (2017). Genome-wide identification and characterization of tRNA-derived RNA fragments in land plants. Plant Mol. Biol. *93*, 35–48.

Ambrožová, K., Mandáková, T., Bureš, P., Neumann, P., Leitch, I.J., Koblížková, A., Macas, J., and Lysak, M.A. (2011). Diverse retrotransposon families and an AT-rich satellite DNA revealed in giant genomes of Fritillaria lilies. Ann. Bot. *107*, 255–268.

Ammar, R., and Thompson, J. (2019). zFPKM: A suite of functions to facilitate zFPKM transformations. R package version 1.8.0, https://github.com/ronammar/zFPKM.

Anders, S., Pyl, P.T., and Huber, W. (2015). HTSeq--a Python framework to work with high-throughput sequencing data. Bioinformatics *31*, 166–169.

Andrews, S. (2010). FastQC A Quality Control tool for High Throughput Sequence Data, http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

Antunez-Sanchez, J., Naish, M., Ramirez-Prado, J.S., Ohno, S., Huang, Y., Dawson, A., Manza-Mianza, D., Ariel, F., Raynaud, C., Wilbowo, A., et al. (2020). A new role for histone demethylases in the maintenance of plant genome integrity. BioRxiv 2020.03.02.972752.

Aronesty, E. (2013). Comparison of Sequencing Utility Programs. Open Bioinforma. J. *7*, 1–8.

Avin-Wittenberg, T. (2019). Autophagy and its role in plant abiotic stress management. Plant Cell Environ. *42*, 1045–1053.

Ayarpadikannan, S., Lee, H.-E., Han, K., and Kim, H.-S. (2015). Transposable element-driven transcript diversification and its relevance to genetic disorders. Gene *558*, 187–194.

Baduel, P., Quadrana, L., Hunter, B., Bomblies, K., and Colot, V. (2019). Relaxed purifying selection in autopolyploids drives transposable element over-accumulation which provides variants for local adaptation. Nat. Commun. *10*, 5818.

Barash, Y., Calarco, J.A., Gao, W., Pan, Q., Wang, X., Shai, O., Blencowe, B.J., and Frey, B.J. (2010). Deciphering the splicing code. Nature *465*, 53–59.

Barkan, A., and Martienssen, R.A. (1991). Inactivation of maize transposon Mu suppresses a mutant phenotype by activating an outward-reading promoter near the end of Mu1. Proc. Natl. Acad. Sci. *88*, 3502–3506.

Bednar, J., Garcia-Saez, I., Boopathi, R., Cutter, A.R., Papai, G., Reymer, A., Syed, S.H., Lone, I.N., Tonchev, O., Crucifix, C., et al. (2017). Structure and Dynamics of a 197 bp Nucleosome in Complex with Linker Histone H1. Mol. Cell *66*, 384-397.e8.

Beguiristain, T., Grandbastien, M.-A., Puigdomènech, P., and Casacuberta, J.M. (2001). Three Tnt1 Subfamilies Show Different Stress-Associated Patterns of Expression in Tobacco. Consequences for Retrotransposon Control and Evolution in Plants. Plant Physiol. *127*, 212–221.

Belyaev, N.D., Houben, A., Baranczewski, P., and Schubert, I. (1997). Histone H4 acetylation in plant heterochromatin is altered during the cell cycle. Chromosoma *106*, 193–197.

Benjak, A., Forneck, A., and Casacuberta, J.M. (2008). Genome-Wide Analysis of the "Cut-and-Paste" Transposons of Grapevine. PLoS ONE *3*, e3107.

Bennetzen, J.L. (2000). Transposable element contributions to plant gene and genome evolution. In Plant Molecular Evolution, J.J. Doyle, and B.S. Gaut, eds. (Dordrecht: Springer Netherlands), pp. 251–269.

Bennetzen, J.L. (2005). Transposable elements, gene creation and genome rearrangement in flowering plants. Curr. Opin. Genet. Dev. *15*, 621–627.

Berger, S.L. (2007). The complex language of chromatin regulation during transcription. Nature *447*, 407–412.

Bissinger, E.-M., Heinke, R., Sippl, W., and Jung, M. (2010). Targeting epigenetic modifiers: Inhibitors of histone methyltransferases. MedChemComm *1*, 114–124.

Blagosklonny, M.V., Robey, R., Sackett, D.L., Du, L., Traganos, F., Darzynkiewicz, Z., Fojo, T., and Bates, S.E. (2002). Histone Deacetylase Inhibitors All Induce p21 but Differentially Cause Tubulin Acetylation, Mitotic Arrest, and Cytotoxicity. Mol. Cancer Ther. *1*, 937–941.

Blagosklonny, M.V., Trostel, S., Kayastha, G., Demidenko, Z.N., Vassilev, L.T., Romanova, L.Y., Bates, S., and Fojo, T. (2005). Depletion of Mutant p53 and Cytotoxicity of Histone Deacetylase Inhibitors. Cancer Res. *65*, 7386–7392.

Bolden, J.E., Peart, M.J., and Johnstone, R.W. (2006). Anticancer activities of histone deacetylase inhibitors. Nat. Rev. Drug Discov. *5*, 769–784.

Bologna, N.G., and Voinnet, O. (2014). The Diversity, Biogenesis, and Activities of Endogenous Silencing Small RNAs in *Arabidopsis*. Annu. Rev. Plant Biol. *65*, 473–503.

Borges, F., and Martienssen, R.A. (2015). The expanding world of small RNAs in plants. Nat. Rev. Mol. Cell Biol. *16*, 727–741.

Borges, F., Parent, J.-S., van Ex, F., Wolff, P., Martínez, G., Köhler, C., and Martienssen, R.A. (2018). Transposon-derived small RNAs triggered by miR845 mediate genome dosage response in Arabidopsis. Nat. Genet. *50*, 186–192.

Bourque, G., Burns, K.H., Gehring, M., Gorbunova, V., Seluanov, A., Hammell, M., Imbeault, M., Izsvák, Z., Levin, H.L., Macfarlan, T.S., et al. (2018). Ten things you should know about transposable elements. Genome Biol. *19*, 199.

Brehm, A., Längst, G., Kehle, J., Clapier, C.R., Imhof, A., Eberharter, A., Müller, J., and Becker, P.B. (2000). dMi-2 and ISWI chromatin remodelling factors have distinct nucleosome binding and mobilization properties. EMBO J. *19*, 4332–4341.

Buchmann, J.P., Löytynoja, A., Wicker, T., and Schulman, A.H. (2014). Analysis of CACTA transposases reveals intron loss as major factor influencing their exon/intron structure in monocotyledonous and eudicotyledonous hosts. Mob. DNA *5*, 24.

Buels, R., Yao, E., Diesh, C.M., Hayes, R.D., Munoz-Torres, M., Helt, G., Goodstein, D.M., Elsik, C.G., Lewis, S.E., Stein, L., et al. (2016). JBrowse: a dynamic web platform for genome visualization and analysis. Genome Biol. *17*, 66.

Butelli, E., Licciardello, C., Zhang, Y., Liu, J., Mackay, S., Bailey, P., Reforgiato-Recupero, G., and Martin, C. (2012). Retrotransposons Control Fruit-Specific, Cold-Dependent Accumulation of Anthocyanins in Blood Oranges. Plant Cell *24*, 1242–1255.

Byrne, A., Beaudin, A.E., Olsen, H.E., Jain, M., Cole, C., Palmer, T., DuBois, R.M., Forsberg, E.C., Akeson, M., and Vollmers, C. (2017). Nanopore long-read RNAseq reveals widespread transcriptional variation among the surface receptors of individual B cells. Nat. Commun. *8*, 16027.

Cantó, C., Gerhart-Hines, Z., Feige, J.N., Lagouge, M., Noriega, L., Milne, J.C., Elliott, P.J., Puigserver, P., and Auwerx, J. (2009). AMPK regulates energy expenditure by modulating NAD + metabolism and SIRT1 activity. Nature *458*, 1056–1060.

Carrier, G., Le Cunff, L., Dereeper, A., Legrand, D., Sabot, F., Bouchez, O., Audeguin, L., Boursiquot, J.-M., and This, P. (2012). Transposable Elements Are a Major Cause of Somatic Polymorphism in Vitis vinifera L. PLoS ONE *7*, e32973.

Casacuberta, E., and González, J. (2013). The impact of transposable elements in environmental adaptation. Mol. Ecol. *22*, 1503–1517.

Cavrak, V.V., Lettner, N., Jamge, S., Kosarewicz, A., Bayer, L.M., and Mittelsten Scheid, O. (2014). How a Retrotransposon Exploits the Plant's Heat Stress Response for Its Activation. PLoS Genet. *10*, e1004115.

Cedillo-Jiménez, C.A., Hernández-Salazar, M., Escobar-Feregrino, T., Caballero-Pérez, J., Arteaga-Vázquez, M., Cruz-Ramírez, A., Torres-Pacheco, I., Guevara-González, R., and Cruz-Hernández, A. (2016). MicroRNAs Sequencing for Understanding the Genetic Regulation of Plant Genomes. In Plant Genomics, I.Y. Abdurakhmonov, ed. (InTech), p.

Chalker-Scott, L. (1999). Environmental Significance of Anthocyanins in Plant Stress Responses. Photochem. Photobiol. *70*, 1–9.

Chang, S., and Pikaard, C.S. (2005). Transcript Profiling in *Arabidopsis* Reveals Complex Responses to Global Inhibition of DNA Methylation and Histone Deacetylation. J. Biol. Chem. *280*, 796–804.

Charlesworth, B. (2009). Effective population size and patterns of molecular evolution and variation. Nat. Rev. Genet. *10*, 195–205.

Chekanova, J.A. (2015). Long non-coding RNAs and their functions in plants. Curr. Opin. Plant Biol. *27*, 207–216.

Chellappan, P., Xia, J., Zhou, X., Gao, S., Zhang, X., Coutino, G., Vazquez, F., Zhang, W., and Jin, H. (2010). siRNAs from miRNA sites mediate DNA methylation of target genes. Nucleic Acids Res. *38*, 6883–6894.

Chen, H., and Boutros, P.C. (2011). VennDiagram: a package for the generation of highly-customizable Venn and Euler diagrams in R. BMC Bioinformatics *12*, 35.

Chen, L.-T., and Wu, K. (2010). Role of histone deacetylases HDA6 and HDA19 in ABA and abiotic stress response. Plant Signal. Behav. *5*, 1318–1320.

Choi, J.Y., and Lee, Y.C.G. (2020). Double-edged sword: The evolutionary consequences of the epigenetic silencing of transposable elements. PLOS Genet. *16*, e1008872.

Cholet, F., Ijaz, U.Z., and Smith, C.J. (2019). Differential ratio amplicons (Ramp) for the evaluation of RNA integrity extracted from complex environmental samples. Environ. Microbiol. *21*, 827–844.

Chow, C.-N., Lee, T.-Y., Hung, Y.-C., Li, G.-Z., Tseng, K.-C., Liu, Y.-H., Kuo, P.-L., Zheng, H.-Q., and Chang, W.-C. (2019). PlantPAN3.0: a new and updated resource for reconstructing transcriptional regulatory networks from ChIP-seq experiments in plants. Nucleic Acids Res. *47*, D1155–D1163.

Chuong, E.B., Elde, N.C., and Feschotte, C. (2016). Regulatory evolution of innate immunity through co-option of endogenous retroviruses. Science *351*, 1083–1087.

Chuong, E.B., Elde, N.C., and Feschotte, C. (2017). Regulatory activities of transposable elements: from conflicts to benefits. Nat. Rev. Genet. *18*, 71–86.

Corona, D.F.V., Siriaco, G., Armstrong, J.A., Snarskaya, N., McClymont, S.A., Scott, M.P., and Tamkun, J.W. (2007). ISWI regulates higher-order chromatin structure and histone H1 assembly in vivo. PLoS Biol. *5*, e232.

Creasey, K.M., Zhai, J., Borges, F., Van Ex, F., Regulski, M., Meyers, B.C., and Martienssen, R.A. (2014). miRNAs trigger widespread epigenetically activated siRNAs from transposons in Arabidopsis. Nature *508*, 411–415.

Cresse, A.D., Hulbert, S.H., Brown, W.E., Lucas, J.R., and Bennetzen, J.L. (1995). Mu1-related transposable elements of maize preferentially insert into low copy number DNA. Genetics *140*, 315–324.

Crevillén, P., Yang, H., Cui, X., Greeff, C., Trick, M., Qiu, Q., Cao, X., and Dean, C. (2014). Epigenetic reprogramming that prevents transgenerational inheritance of the vernalized state. Nature *515*, 587–590.

Cuerda-Gil, D., and Slotkin, R.K. (2016). Non-canonical RNA-directed DNA methylation. Nat. Plants *2*, 16163.

Cuperus, J.T., Fahlgren, N., and Carrington, J.C. (2011). Evolution and Functional Diversification of MIRNA Genes. Plant Cell *23*, 431–442.

Czech, B., and Hannon, G.J. (2016). One Loop to Rule Them All: The Ping-Pong Cycle and piRNA-Guided Silencing. Trends Biochem. Sci. *41*, 324–337.

Dai, X., and Zhao, P.X. (2011). psRNATarget: a plant small RNA target analysis server. Nucleic Acids Res. *39*, W155–W159.

Dangl, M., Brosch, G., Haas, H., Loidl, P., and Lusser, A. (2001). Comparative analysis of HD2 type histone deacetylases in higher plants. Planta *213*, 280–285.

De Coster, W., D'Hert, S., Schultz, D.T., Cruts, M., and Van Broeckhoven, C. (2018). NanoPack: visualizing and processing long-read sequencing data. Bioinformatics *34*, 2666–2669.

Demidchik, V. (2015). Mechanisms of oxidative stress in plants: From classical chemistry to cell biology. Environ. Exp. Bot. *109*, 212–228.

Deniz, Ö., Frost, J.M., and Branco, M.R. (2019). Regulation of transposable elements by DNA modifications. Nat. Rev. Genet. *20*, 417–431.

Deragon, J.-M., and Zhang, X. (2006). Short Interspersed Elements (SINEs) in Plants: Origin, Classification, and Use as Phylogenetic Markers. Syst. Biol. *55*, 949–956.

Deremetz, A., Roux, C.L., Idir, Y., Brousse, C., Agorio, A., Gy, I., Parker, J.E., and Bouché, N. (2019). Antagonistic Actions of FPA and IBM2 Regulate Transcript Processing from Genes Containing Heterochromatin. Plant Physiol. *180*, 392–403.

Devoto, A., Nieto-Rostro, M., Xie, D., Ellis, C., Harmston, R., Patrick, E., Davis, J., Sherratt, L., Coleman, M., and Turner, J.G. (2002). COI1 links jasmonate signalling and fertility to the SCF ubiquitin–ligase complex in Arabidopsis. Plant J. *32*, 457–466.

Di, C., Yuan, J., Wu, Y., Li, J., Lin, H., Hu, L., Zhang, T., Qi, Y., Gerstein, M.B., Guo, Y., et al. (2014). Characterization of stress-responsive lncRNAs in Arabidopsis thaliana by integrating expression, epigenetic and structural features. Plant J. *80*, 848–861.

Díaz-Riquelme, J., Zhurov, V., Rioja, C., Pérez-Moreno, I., Torres-Pérez, R., Grimplet, J., Carbonell-Bejerano, P., Bajda, S., Van Leeuwen, T., Martínez-Zapater, J.M., et al. (2016). Comparative genome-wide transcriptome analysis of Vitis vinifera responses to adapted and non-adapted strains of two-spotted spider mite, Tetranyhus urticae. BMC Genomics *17*, 74.

van Dijk, E.L., Auger, H., Jaszczyszyn, Y., and Thermes, C. (2014). Ten years of next-generation sequencing technology. Trends Genet. *30*, 418–426.

Dowen, R.H., Pelizzola, M., Schmitz, R.J., Lister, R., Dowen, J.M., Nery, J.R., Dixon, J.E., and Ecker, J.R. (2012). Widespread dynamic DNA methylation in response to biotic stress. Proc. Natl. Acad. Sci. *109*, E2183–E2191.

Earley, K., Lawrence, R.J., Pontes, O., Reuther, R., Enciso, A.J., Silva, M., Neves, N., Gross, M., Viegas, W., and Pikaard, C.S. (2006). Erasure of histone acetylation by Arabidopsis HDA6 mediates large-scale gene silencing in nucleolar dominance. Genes Dev. *20*, 1283–1293.

Earley, K.W., Pontvianne, F., Wierzbicki, A.T., Blevins, T., Tucker, S., Costa-Nunes, P., Pontes, O., and Pikaard, C.S. (2010). Mechanisms of HDA6-mediated rRNA gene silencing: suppression of intergenic Pol II transcription and differential effects on maintenance versus siRNA-directed cytosine methylation. Genes Dev. *24*, 1119–1132.

Ebbs, M.L., and Bender, J. (2006). Locus-Specific Control of DNA Methylation by the Arabidopsis SUVH5 Histone Methyltransferase. Plant Cell *18*, 1166–1176.

Edgar, R.C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. *32*, 1792–1797.

Erice, A., Brambilla, D., Bremer, J., Jackson, J.B., Kokka, R., Yen-Lieberman, B., and Coombs, R.W. (2000). Performance Characteristics of the QUANTIPLEX HIV-1 RNA 3.0 Assay for Detection and Quantitation of Human Immunodeficiency Virus Type 1 RNA in Plasma. J. Clin. Microbiol. *38*, 2837–2845.

Falkenberg, K.J., and Johnstone, R.W. (2014). Histone deacetylases and their inhibitors in cancer, neurological diseases and immune disorders. Nat. Rev. Drug Discov. *13*, 673–691.

Fedoroff, N.V. (2012). Transposable Elements, Epigenetics, and Genome Evolution. Science *338*, 758–767.

Ferrer, M., Henriet, S., Chamontin, C., Lainé, S., and Mougel, M. (2016). From Cells to Virus Particles: Quantitative Methods to Monitor RNA Packaging. Viruses *8*.

Feschotte, C. (2008). Transposable elements and the evolution of regulatory networks. Nat. Rev. Genet. *9*, 397–405.

Finnegan, D.J. (1992). Transposable elements. Curr. Opin. Genet. Dev. *2*, 861–867.

Flematti, G.R., Dixon, K.W., and Smith, S.M. (2015). What are karrikins and how were they 'discovered' by plants? BMC Biol. *13*, 108.

Flemr, M., Malik, R., Franke, V., Nejepinska, J., Sedlacek, R., Vlahovicek, K., and Svoboda, P. (2013). A Retrotransposon-Driven Dicer Isoform Directs Endogenous Small Interfering RNA Production in Mouse Oocytes. Cell *155*, 807–816.

Foster, T.M., and Aranzana, M.J. (2018). Attention sports fans! The far-reaching contributions of bud sport mutants to horticulture and plant biology. Hortic. Res. *5*, 44.

Frey, M., Reinecke, J., Grant, S., Saedler, H., and Gierl, A. (1990). Excision of the En/Spm transposable element of Zea mays requires two element-encoded proteins. EMBO J. *9*, 4037–4044.

Fultz, D., Choudury, S.G., and Slotkin, R.K. (2015). Silencing of active transposable elements in plants. Curr. Opin. Plant Biol. *27*, 67–76.

Gao, L., Cueto, M.A., Asselbergs, F., and Atadja, P. (2002). Cloning and Functional Characterization of HDAC11, a Novel Member of the Human Histone Deacetylase Family. J. Biol. Chem. *277*, 25748–25755.

Garalde, D.R., Snell, E.A., Jachimowicz, D., Sipos, B., Lloyd, J.H., Bruce, M., Pantic, N., Admassu, T., James, P., Warland, A., et al. (2018). Highly parallel direct RNA sequencing on an array of nanopores. Nat. Methods *15*, 201–206.

Gaut, B.S., Morton, B.R., McCaig, B.C., and Clegg, M.T. (1996). Substitution rate comparisons between grasses and palms: synonymous rate differences at the nuclear gene Adh parallel rate differences at the plastid gene rbcL. Proc. Natl. Acad. Sci. *93*, 10274–10279.

Gendrel, A.-V. (2002). Dependence of Heterochromatic Histone H3 Methylation Patterns on the Arabidopsis Gene DDM1. Science *297*, 1871–1873.

Girard, A., and Hannon, G.J. (2008). Conserved themes in small-RNA-mediated transposon control. Trends Cell Biol. *18*, 136–148.

Goh, C.-H., Ko, S.-M., Koh, S., Kim, Y.-J., and Bae, H.-J. (2012). Photosynthesis and Environments: Photoinhibition and Repair Mechanisms in Plants. J. Plant Biol. *55*, 93–101.

Goodwin, S., Gurtowski, J., Ethe-Sayers, S., Deshpande, P., Schatz, M.C., and McCombie, W.R. (2015). Oxford Nanopore sequencing, hybrid error correction, and de novo assembly of a eukaryotic genome. Genome Res. *25*, 1750–1756.

Gore, S.D., and Carducci, M.A. (2000). Modifying histones to tame cancer: clinical development of sodium phenylbutyrate and other histone deacetylase inhibitors. Expert Opin. Investig. Drugs *9*, 2923–2934.

Grabundzija, I., Messing, S.A., Thomas, J., Cosby, R.L., Bilic, I., Miskey, C., Gogol-Döring, A., Kapitonov, V., Diem, T., Dalda, A., et al. (2016). A Helitron transposon reconstructed from bats reveals a novel mechanism of genome shuffling in eukaryotes. Nat. Commun. *7*, 10716.

Grandbastien, M.-A. (1998). Activation of plant retrotransposons under stress conditions. Trends Plant Sci. *3*, 181–187.

Greenberg, M.V.C., Deleris, A., Hale, C.J., Liu, A., Feng, S., and Jacobsen, S.E. (2013). Interplay between Active Chromatin Marks and RNA-Directed DNA Methylation in Arabidopsis thaliana. PLoS Genet. *9*, e1003946.

Grzebelus, D., Lasota, S., Gambin, T., Kucherov, G., and Gambin, A. (2007). Diversity and structure of PIF/Harbinger-like elements in the genome of Medicago truncatula. BMC Genomics *8*, 409.

Guerreiro, M.P.G. (2012). What makes transposable elements move in the Drosophila genome? Heredity *108*, 461–468.

Hannah, L., Roehrdanz, P.R., Ikegami, M., Shepard, A.V., Shaw, M.R., Tabor, G., Zhi, L., Marquet, P.A., and Hijmans, R.J. (2013). Climate change, wine, and conservation. Proc. Natl. Acad. Sci. *110*, 6907–6912.

Harold Pimentel (2014). What the FPKM? A review of RNA-Seq expression units.

Hart, T., Komori, H., LaMere, S., Podshivalova, K., and Salomon, D.R. (2013). Finding the active genes in deep RNA-seq gene expression studies. BMC Genomics *14*, 778.

Hashida, S.-N., Uchiyama, T., Martin, C., Kishima, Y., Sano, Y., and Mikami, T. (2006). The Temperature-Dependent Change in Methylation of the *Antirrhinum* Transposon Tam3 Is Controlled by the Activity of Its Transposase. Plant Cell *18*, 104–118.

Haussecker, D., Huang, Y., Lau, A., Parameswaran, P., Fire, A.Z., and Kay, M.A. (2010). Human tRNA-derived small RNAs in the global regulation of RNA silencing. RNA *16*, 673–695.

Hayashi, K., and Yoshida, H. (2009). Refunctionalization of the ancient rice blast disease resistance gene *Pit* by the recruitment of a retrotransposon as a promoter. Plant J. *57*, 413–425.

Hayashi, S., and Takaiwa, F. (2015). Visualization of endoplasmic reticulum stressed cells for forward genetic studies in plants. J. Plant Physiol. *180*, 61–66.

He, F., and Jacobson, A. (2015). Nonsense-Mediated mRNA Decay: Degradation of Defective Transcripts Is Only Part of the Story. Annu. Rev. Genet. *49*, 339–366.

Hirochika, H. (1993). Activation of tobacco retrotransposons during tissue culture. EMBO J. *12*, 2521–2528.

Hirochika, H., Sugimoto, K., Otsuki, Y., Tsugawa, H., and Kanda, M. (1996). Retrotransposons of rice involved in mutations induced by tissue culture. Proc. Natl. Acad. Sci. *93*, 7783–7788.

Hirochika, H., Okamoto, H., and Kakutani, T. (2000). Silencing of Retrotransposons in Arabidopsis and Reactivation by the ddm1 Mutation. Plant Cell *12*, 357–368.

Hirsch, C.D., and Springer, N.M. (2017). Transposable element influences on gene expression in plants. Biochim. Biophys. Acta BBA - Gene Regul. Mech. *1860*, 157–165.

Hnilicová, J., Hozeifi, S., Dušková, E., Icha, J., Tománková, T., and Staněk, D. (2011). Histone Deacetylase Activity Modulates Alternative Splicing. PLOS ONE *6*, e16727.

Hoat, T.X., Nakayashiki, H., Tosa, Y., and Mayama, S. (2006). Specific cleavage of ribosomal RNA and mRNA during victorin-induced apoptotic cell death in oat. Plant J. *46*, 922–933.

Hollister, J.D., and Gaut, B.S. (2009). Epigenetic silencing of transposable elements: A trade-off between reduced transposition and deleterious effects on neighboring gene expression. Genome Res. *19*, 1419–1428.

Honda, S., Kawamura, T., Loher, P., Morichika, K., Rigoutsos, I., and Kirino, Y. (2017). The biogenesis pathway of tRNA-derived piRNAs in Bombyx germ cells. Nucleic Acids Res. *45*, 9108–9120.

Hu, Y., Zhang, L., He, S., Huang, M., Tan, J., Zhao, L., Yan, S., Li, H., Zhou, K., Liang, Y., et al. (2012). Cold stress selectively unsilences tandem repeats in heterochromatin associated with accumulation of H3K9ac. Plant Cell Environ. *35*, 2130–2142.

Huang, C.R.L., Burns, K.H., and Boeke, J.D. (2012). Active Transposition in Genomes. Annu. Rev. Genet. *46*, 651–675.

Huang, X., Fejes Tóth, K., and Aravin, A.A. (2017). piRNA Biogenesis in Drosophila melanogaster. Trends Genet. *33*, 882–894.

Hyndman, R. (2018). hdrcde: Highest Density Regions and Conditional Density Estimation. R package version 3.3.

Iannitti, T., and Palmieri, B. (2011). Clinical and Experimental Applications of Sodium Phenylbutyrate. Drugs R D 23.

Ibarra-Laclette, E., Lyons, E., Hernández-Guzmán, G., Pérez-Torres, C.A., Carretero-Paulet, L., Chang, T.-H., Lan, T., Welch, A.J., Juárez, M.J.A., Simpson, J., et al. (2013). Architecture and evolution of a minute plant genome. Nature *498*, 94–98.

Imai, S., Armstrong, C.M., Kaeberlein, M., and Guarente, L. (2000). Transcriptional silencing and longevity protein Sir2 is an NAD-dependent histone deacetylase. Nature *403*, 795–800.

Inagaki, Y., Hisatomi, Y., Suzuki, T., Kasahara, K., and Iida, S. (1994). Isolation of a Suppressor-mutator/Enhancer-like transposable element, Tpn1, from Japanese morning glory bearing variegated flowers. Plant Cell *6*, 375–383.

International Human Genome Sequencing Consortium (2001). Initial sequencing and analysis of the human genome. Nature *409*, 860–921.

International Rice Genome Sequencing Project, and Sasaki, T. (2005). The map-based sequence of the rice genome. Nature *436*, 793–800.

Ito, H., Gaubert, H., Bucher, E., Mirouze, M., Vaillant, I., and Paszkowski, J. (2011). An siRNA pathway prevents transgenerational retrotransposition in plants subjected to stress. Nature *472*, 115–119.

Ito, H., Kim, J.-M., Matsunaga, W., Saze, H., Matsui, A., Endo, T.A., Harukawa, Y., Takagi, H., Yaegashi, H., Masuta, Y., et al. (2016). A Stress-Activated Transposon in Arabidopsis Induces Transgenerational Abscisic Acid Insensitivity. Sci. Rep. *6*, 23181.

Jabre, I., Reddy, A.S.N., Kalyna, M., Chaudhary, S., Khokhar, W., Byrne, L.J., Wilson, C.M., and Syed, N.H. (2019). Does co-transcriptional regulation of alternative splicing mediate plant stress responses? Nucleic Acids Res. *47*, 2716–2726.

Jackson, J.P., Johnson, L., Jasencakova, Z., Zhang, X., PerezBurgos, L., Singh, P.B., Cheng, X., Schubert, I., Jenuwein, T., and Jacobsen, S.E. (2004). Dimethylation of histone H3 lysine 9 is a critical mark for DNA methylation and gene silencing in Arabidopsis thaliana. Chromosoma *112*, 308–315.

Jaillon, O., Aury, J.-M., Noel, B., Policriti, A., Clepet, C., Casagrande, A., Choisne, N., Aubourg, S., Vitulo, N., Jubin, C., et al. (2007). The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. Nature *449*, 463–467.

Jain, M., Tyson, J.R., Loose, M., Ip, C.L.C., Eccles, D.A., O'Grady, J., Malla, S., Leggett, R.M., Wallerman, O., Jansen, H.J., et al. (2017). MinION Analysis and Reference Consortium: Phase 2 data release and analysis of R9.0 chemistry. F1000Research *6*.

Jang, I.-C., Chung, P.J., Hemmes, H., Jung, C., and Chua, N.-H. (2011). Rapid and Reversible Light-Mediated Chromatin Modifications of Arabidopsis Phytochrome A Locus[C][W]. Plant Cell *23*, 459–470.

Jedlicka, P., Lexa, M., Vanat, I., Hobza, R., and Kejnovsky, E. (2019). Nested plant LTR retrotransposons target specific regions of other elements, while all LTR retrotransposons often target palindromes and nucleosome-occupied regions: in silico study. Mob. DNA *10*, 50.

Jiang, N., Bao, Z., Zhang, X., Hirochika, H., Eddy, S.R., McCouch, S.R., and Wessler, S.R. (2003). An active DNA transposon family in rice. Nature *421*, 163–167.

Jin, Y., Tam, O.H., Paniagua, E., and Hammell, M. (2015). TEtranscripts: a package for including transposable elements in differential expression analysis of RNA-seq datasets. Bioinformatics *31*, 3593–3599.

Johnson, L.M., Bostick, M., Zhang, X., Kraft, E., Henderson, I., Callis, J., and Jacobsen, S.E. (2007). The SRA Methyl-Cytosine-Binding Domain Links DNA and Histone Methylation. Curr. Biol. *17*, 379–384.

Jordan, I.K., Rogozin, I.B., Glazko, G.V., and Koonin, E.V. (2003). Origin of a substantial fraction of human regulatory sequences from transposable elements. Trends Genet. *19*, 68–72.

Jukes, T.H., and Cantor, C.R. (1969). Evolution of Protein Molecules. New York: Academic Press (New York: Academic Press).

Jung, H., Winefield, C., Bombarely, A., Prentis, P., and Waterhouse, P. (2019). Tools and Strategies for Long-Read Sequencing and De Novo Assembly of Plant Genomes. Trends Plant Sci. *24*, 700–724.

Kabelitz, T., Kappel, C., Henneberger, K., Benke, E., Nöh, C., and Bäurle, I. (2014). eQTL Mapping of Transposon Silencing Reveals a Position-Dependent Stable Escape from Epigenetic Silencing and Transposition of AtMu1 in the Arabidopsis Lineage. Plant Cell *26*, 3261–3271.

Kapitonov, V.V., and Jurka, J. (2007). Helitrons on a roll: eukaryotic rolling-circle transposons. Trends Genet. *23*, 521–529.

Keam, S.P., Young, P.E., McCorkindale, A.L., Dang, T.H.Y., Clancy, J.L., Humphreys, D.T., Preiss, T., Hutvagner, G., Martin, D.I.K., Cropley, J.E., et al. (2014). The human Piwi protein Hiwi2 associates with tRNA-derived piRNAs in somatic cells. Nucleic Acids Res. *42*, 8984–8995.

Keller, E.F. (1983). A Feeling for the Organism: The Life and Work of Barbara McClintock (Henry Holt and Company).

Khraiwesh, B., Arif, M.A., Seumel, G.I., Ossowski, S., Weigel, D., Reski, R., and Frank, W. (2010). Transcriptional Control of Gene Expression by MicroRNAs. Cell *140*, 111–122.

Kilburn, D., Burke, J., Fedak, R., Olsen, H., Jain, M., Miga, K., Mayes, S., and Liu, K. (2020). High Data Throughput and Low Cost Ultra Long Nanopore Sequencing.

Kim, D., Langmead, B., and Salzberg, S.L. (2015a). HISAT: a fast spliced aligner with low memory requirements. Nat. Methods *12*, 357–360.

Kim, K.D., Baidouri, M.E., Abernathy, B., Iwata-Otsubo, A., Chavarro, C., Gonzales, M., Libault, M., Grimwood, J., and Jackson, S.A. (2015b). A Comparative Epigenomic Analysis of Polyploidy-Derived Genes in Soybean and Common Bean. Plant Physiol. *168*, 1433–1447.

Knip, M., de Pater, S., and Hooykaas, P.J. (2012). The SLEEPER genes: a transposase-derived angiosperm-specific gene family. BMC Plant Biol. *12*, 192.

Krug, L., Chatterjee, N., Borges-Monroy, R., Hearn, S., Liao, W.-W., Morrill, K., Prazak, L., Rozhkov, N., Theodorou, D., Hammell, M., et al. (2017). Retrotransposon activation contributes to neurodegeneration in a Drosophila TDP-43 model of ALS. PLOS Genet. *13*, e1006635.

Kuang, H., Padmanabhan, C., Li, F., Kamei, A., Bhaskar, P.B., Ouyang, S., Jiang, J., Buell, C.R., and Baker, B. (2009). Identification of miniature inverted-repeat transposable elements (MITEs) and biogenesis of their siRNAs in the Solanaceae: New functional implications for MITEs. Genome Res. *19*, 42–56.

Kubicek, S., O'Sullivan, R.J., August, E.M., Hickey, E.R., Zhang, Q., Teodoro, M.L., Rea, S., Mechtler, K., Kowalski, J.A., Homon, C.A., et al. (2007). Reversal of H3K9me2 by a Small-Molecule Inhibitor for the G9a Histone Methyltransferase. Mol. Cell *25*, 473–481.

Kubota, K., Niinuma, Y., Kaneko, M., Okuma, Y., Sugai, M., Omura, T., Uesugi, M., Uehara, T., Hosoi, T., and Nomura, Y. (2006). Suppressive effects of 4-phenylbutyrate on the aggregation of Pael receptors and endoplasmic reticulum stress. J. Neurochem. *97*, 1259–1268.

Kullan, J.B., Pinto, D.L.P., Bertolini, E., Fasoli, M., Zenoni, S., Tornielli, G.B., Pezzotti, M., Meyers, B.C., Farina, L., Pè, M.E., et al. (2015). miRVine: a microRNA expression atlas of grapevine based on small RNA sequencing. BMC Genomics *16*, 393.

Kumar, P., Anaya, J., Mudunuri, S.B., and Dutta, A. (2014). Meta-analysis of tRNA derived RNA fragments reveals that they are evolutionarily conserved and associate with AGO proteins to recognize specific RNA targets. BMC Biol. *12*, 78.

Kurdistani, S.K., and Grunstein, M. (2003). Histone acetylation and deacetylation in yeast. Nat. Rev. Mol. Cell Biol. *4*, 276–284.

Kurdistani, S.K., Robyr, D., Tavazoie, S., and Grunstein, M. (2002). Genome-wide binding map of the histone deacetylase Rpd3 in yeast. Nat. Genet. N. Y. *31*, 248–254.

Kurita, K., Sakamoto, T., Yagi, N., Sakamoto, Y., Ito, A., Nishino, N., Sako, K., Yoshida, M., Kimura, H., Seki, M., et al. (2017). Live imaging of H3K9 acetylation in plant cells. Sci. Rep. *7*, 45894.

Kusaczuk, M., Bartoszewicz, M., and Cechowska-Pasko, M. (2015). Phenylbutyric Acid: Simple Structure - Multiple Effects. Curr. Pharm. Des. *21*, 2147–2166.

Kusaczuk, M., Krętowski, R., Bartoszewicz, M., and Cechowska-Pasko, M. (2016). Phenylbutyrate—a pan-HDAC inhibitor—suppresses proliferation of glioblastoma LN-229 cell line. Tumor Biol. *37*, 931–942.

Lanciano, S., and Cristofari, G. (2020). Measuring and interpreting transposable element expression. Nat. Rev. Genet. *21*, 721–736.

Langmead, B. (2010). Aligning short sequencing reads with Bowtie. Curr. Protoc. Bioinforma. Ed. Board Andreas Baxevanis Al *CHAPTER*, Unit-11.7.

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S.L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. *10*, R25.

Law, J.A., and Jacobsen, S.E. (2010). Establishing, maintaining and modifying DNA methylation patterns in plants and animals. Nat. Rev. Genet. Lond. *11*, 204–220.

Law, R.D., and Suttle, J.C. (2005). Chromatin remodeling in plant cell culture: patterns of DNA methylation and histone H3 and H4 acetylation vary during growth of asynchronous potato cell suspensions. Plant Physiol. Biochem. *43*, 527–534.

Law, J.A., Du, J., Hale, C.J., Feng, S., Krajewski, K., Palanca, A.M.S., Strahl, B.D., Patel, D.J., and Jacobsen, S.E. (2013). Polymerase IV occupancy at RNA-directed DNA methylation sites requires SHH1. Nature *498*, 385–389.

Le, T.N., Miyazaki, Y., Takuno, S., and Saze, H. (2015). Epigenetic regulation of intragenic transposable elements impacts gene transcription in Arabidopsis thaliana. Nucleic Acids Res. *43*, 3911–3921.

Lee, C., Grasso, C., and Sharlow, M.F. (2002). Multiple sequence alignment using partial order graphs. Bioinformatics *18*, 452–464.

Lee, S.C., Ernst, E., Berube, B., Borges, F., Parent, J.-S., Ledon, P., Schorn, A., and Martienssen, R.A. (2020). Arabidopsis retrotransposon virus-like particles and their regulation by epigenetically activated small RNA. Genome Res. *30*, 576–588.

Lefevere, H., Bauters, L., and Gheysen, G. (2020). Salicylic Acid Biosynthesis in Plants. Front. Plant Sci. *11*.

Lei, M., La, H., Lu, K., Wang, P., Miki, D., Ren, Z., Duan, C.-G., Wang, X., Tang, K., Zeng, L., et al. (2014). Arabidopsis EDM2 promotes IBM1 distal polyadenylation and regulates genome DNA methylation patterns. Proc. Natl. Acad. Sci. *111*, 527–532.

Leng, X., Fang, J., Pervaiz, T., Li, Y., Wang, X., Liu, D., Zhu, X., and Fang, J. (2015). Characterization of Expression Patterns of Grapevine MicroRNA Family Members using MicroRNA Rapid Amplification of Complementary DNA Ends. Plant Genome *8*, plantgenome2014.10.0069.

Lev Maor, G., Yearim, A., and Ast, G. (2015). The alternative role of DNA methylation in splicing regulation. Trends Genet. TIG *31*, 274–280.

Lev-Maor, G., Ram, O., Kim, E., Sela, N., Goren, A., Levanon, E.Y., and Ast, G. (2008). Intronic Alus Influence Alternative Splicing. PLoS Genet. *4*, e1000204.

Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics *34*, 3094–3100.

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics *25*, 1754–1760.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078–2079.

Li, H., Soriano, M., Cordewener, J., Muiño, J.M., Riksen, T., Fukuoka, H., Angenent, G.C., and Boutilier, K. (2014). The Histone Deacetylase Inhibitor Trichostatin A Promotes Totipotency in the Male Gametophyte. Plant Cell *26*, 195–209.

Li, J., Akagi, K., Hu, Y., Trivett, A.L., Hlynialuk, C.J.W., Swing, D.A., Volfovsky, N., Morgan, T.C., Golubeva, Y., Stephens, R.M., et al. (2012a). Mouse endogenous retroviruses can trigger premature transcriptional termination at a distance. Genome Res. *22*, 870–884.

Li, S., Vandivier, L.E., Tu, B., Gao, L., Won, S.Y., Li, S., Zheng, B., Gregory, B.D., and Chen, X. (2015). Detection of Pol IV/RDR2-dependent transcripts at the genomic scale in Arabidopsis reveals features and regulation of siRNA biogenesis. Genome Res. *25*, 235–245.

Li, X., Qian, W., Zhao, Y., Wang, C., Shen, J., Zhu, J.-K., and Gong, Z. (2012b). Antisilencing role of the RNA-directed DNA methylation pathway and a histone acetyltransferase in Arabidopsis. Proc. Natl. Acad. Sci. *109*, 11425–11430.

Li, Y., Butenko, Y., and Grafi, G. (2005). Histone deacetylation is required for progression through mitosis in tobacco cells. Plant J. *41*, 346–352.

Liang, Z., Duan, S., Sheng, J., Zhu, S., Ni, X., Shao, J., Liu, C., Nick, P., Du, F., Fan, P., et al. (2019). Whole-genome resequencing of 472 Vitis accessions for grapevine diversity and demographic history analyses. Nat. Commun. *10*, 1190.

Lindroth, A.M., Cao, X., Jackson, J.P., Zilberman, D., McCallum, C.M., Henikoff, S., and Jacobsen, S.E. (2001). Requirement of CHROMOMETHYLASE3 for Maintenance of CpXpG Methylation. Science *292*, 2077–2080.

Lisch, D. (2013). How important are transposons for plant evolution? Nat. Rev. Genet. *14*, 49–61.

Liu, Q., Wang, X., Yang, R., Yang, L., Sun, B., and Zhu, L. (2019). Uptake Kinetics, Accumulation, and Long-Distance Transport of Organophosphate Esters in Plants: Impacts of Chemical and Plant Properties. Environ. Sci. Technol. *53*, 4940–4947.

Liu, S., Yeh, C.-T., Ji, T., Ying, K., Wu, H., Tang, H.M., Fu, Y., Nettleton, D., and Schnable, P.S. (2009). Mu Transposon Insertion Sites and Meiotic Recombination Events Co-Localize with Epigenetic Marks for Open Chromatin across the Maize Genome. PLOS Genet. *5*, e1000733.

Liu, X., Yu, C.-W., Duan, J., Luo, M., Wang, K., Tian, G., Cui, Y., and Wu, K. (2012). HDA6 Directly Interacts with DNA Methyltransferase MET1 and Maintains Transposable Element Silencing in Arabidopsis. Plant Physiol. *158*, 119–129.

Liu, Z., Yan, J.-P., Li, D.-K., Luo, Q., Yan, Q., Liu, Z.-B., Ye, L.-M., Wang, J.-M., Li, X.-F., and Yang, Y. (2015). UDP-Glucosyltransferase71C5, a Major Glucosyltransferase, Mediates Abscisic Acid Homeostasis in Arabidopsis1[OPEN]. Plant Physiol. *167*, 1659–1670.

Lizamore, D.K. (2013). A study of endogenous transposon activity in grapevine (Vitis vinifera L.). Lincoln university.

Lizamore, D., and Winefield, C. (2017). A comparative survey of small RNA and their targets in grapevine embryogenic callus cultures and young leaves. Acta Hortic. 329–336.

Lock, F.E., Rebollo, R., Miceli-Royer, K., Gagnier, L., Kuah, S., Babaian, A., Sistiaga-Poveda, M., Lai, C.B., Nemirovsky, O., Serrano, I., et al. (2014). Distinct isoform of FABP7 revealed by screening for retroelement-activated genes in diffuse large B-cell lymphoma. Proc. Natl. Acad. Sci. *111*, E3534–E3543.

Lorenz, D.A., Sathe, S., Einstein, J.M., and Yeo, G.W. (2020). Direct RNA sequencing enables m6A detection in endogenous transcript isoforms at base-specific resolution. RNA *26*, 19–28.

Love, M.I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. *15*, 550.

Luan, S. (2002). Tyrosine phosphorylation in plant cell signaling. Proc. Natl. Acad. Sci. *99*, 11567–11569.

Lusser, A., Brosch, G., Loidl, A., Haas, H., and Loidl, P. (1997). Identification of Maize Histone Deacetylase HD2 as an Acidic Nucleolar Phosphoprotein. Science *277*, 88–91.

Lusser, M., Parisi, C., Plan, D., and Rodríguez-Cerezo, E. (2012). Deployment of new biotechnologies in plant breeding. Nat. Biotechnol. *30*, 231–239.

Ma, X., Lv, S., Zhang, C., and Yang, C. (2013). Histone deacetylases and their functions in plants. Plant Cell Rep. *32*, 465–478.

Makarevitch, I., Waters, A.J., West, P.T., Stitzer, M., Hirsch, C.N., Ross-Ibarra, J., and Springer, N.M. (2015). Transposable Elements Contribute to Activation of Maize Genes in Response to Abiotic Stress. PLoS Genet. *11*, e1004915.

Malheiro, A., Santos, J., Fraga, H., and Pinto, J. (2010). Climate change scenarios applied to viticultural zoning in Europe. Clim. Res. *43*, 163–177.

Małolepszy, A., Mun, T., Sandal, N., Gupta, V., Dubin, M., Urbański, D., Shah, N., Bachmann, A., Fukai, E., Hirakawa, H., et al. (2016). The *LORE1* insertion mutant resource. Plant J. *88*, 306–317.

Mao, L., Wood, T.C., Yu, Y., Budiman, M.A., Tomkins, J., Woo, S., Sasinowski, M., Presting, G., Frisch, D., Goff, S., et al. (2000). Rice Transposable Elements: A Survey of 73,000 Sequence-Tagged-Connectors. Genome Res. *10*, 982–990.

Marí-Ordóñez, A., Marchais, A., Etcheverry, M., Martin, A., Colot, V., and Voinnet, O. (2013). Reconstructing de novo silencing of an active plant retrotransposon. Nat. Genet. *45*, 1029–1039.

Martin, S.L. (2010). Nucleic acid chaperone properties of ORF1p from the non-LTR retrotransposon, LINE-1. RNA Biol. *7*, 706–711.

Martínez, G., and Slotkin, R.K. (2012). Developmental relaxation of transposable element silencing in plants: functional or byproduct? Curr. Opin. Plant Biol. *15*, 496–502.

Martinez, G., Choudury, S.G., and Slotkin, R.K. (2017). tRNA-derived small RNAs target transposable element transcripts. Nucleic Acids Res. *45*, 5142–5152.

Mathieu, O., Reinders, J., Čaikovski, M., Smathajitt, C., and Paszkowski, J. (2007). Transgenerational Stability of the Arabidopsis Epigenome Is Coordinated by CG Methylation. Cell *130*, 851–862.

McAnena, P., Brown, J.A.L., and Kerin, M.J. (2017). Circulating Nucleosomes and Nucleosome Modifications as Biomarkers in Cancer. Cancers *9*, 5.

McClintock, B. (1950). The origin and behavior of mutable loci in maize. Proc. Natl. Acad. Sci. *36*, 344–355.

McConnell, M.J., Moran, J.V., Abyzov, A., Akbarian, S., Bae, T., Cortes-Ciriano, I., Erwin, J.A., Fasching, L., Flasch, D.A., Freed, D., et al. (2017). Intersection of diverse neuronal genomes and neuropsychiatric disease: The Brain Somatic Mosaicism Network. Science *356*, eaal1641.

McCue, A.D., Panda, K., Nuthikattu, S., Choudury, S.G., Thomas, E.N., and Slotkin, R.K. (2015). ARGONAUTE 6 bridges transposable element mRNA-derived siRNAs to the establishment of DNA methylation. EMBO J. *34*, 20–35.

Mejía-Guerra, M.K., Li, W., Galeano, N.F., Vidal, M., Gray, J., Doseff, A.I., and Grotewold, E. (2015). Core Promoter Plasticity Between Maize Tissues and Genotypes Contrasts with Predominance of Sharp Transcription Initiation Sites. Plant Cell *27*, 3309–3320.

Mérel, V., Boulesteix, M., Fablet, M., and Vieira, C. (2020). Transposable elements in Drosophila. Mob. DNA *11*, 23.

Metcalfe, C.J., and Casane, D. (2013). Accommodating the load. Mob. Genet. Elem. *3*, e24775.

Minucci, S., and Pelicci, P.G. (2006). Histone deacetylase inhibitors and the promise of epigenetic (and more) treatments for cancer. Nat. Rev. Cancer *6*, 38–51.

Misteli, T., Gunjan, A., Hock, R., Bustin, M., and Brown, D.T. (2000). Dynamic binding of histone H1 to chromatin in living cells. Nat. Lond. *408*, 877–881.

Miura, A., Yonebayashi, S., Watanabe, K., Toyama, T., Shimada, H., and Kakutani, T. (2001). Mobilization of transposons by a mutation abolishing full DNA methylation in Arabidopsis. Nature *411*, 212–214.

Miyao, A., Tanaka, K., Murata, K., Sawaki, H., Takeda, S., Abe, K., Shinozuka, Y., Onosato, K., and Hirochika, H. (2003). Target Site Specificity of the *Tos17* Retrotransposon Shows a Preference for Insertion within Genes and against Insertion in Retrotransposon-Rich Regions of the Genome. Plant Cell *15*, 1771–1780.

Moisy, C., Garrison, K., Meredith, C.P., and Pelsy, F. (2008). Characterization of ten novel Ty1 copia-like retrotransposon families of the grapevine genome. BMC Genomics *9*, 469.

Moore, M.J. (2005). From Birth to Death: The Complex Lives of Eukaryotic mRNAs. Science *309*, 1514–1518.

Mottamal, M., Zheng, S., Huang, T.L., and Wang, G. (2015). Histone Deacetylase Inhibitors in Clinical Studies as Templates for New Anticancer Agents. Molecules *20*, 3898–3941.

Mouse Genome Sequencing Consortium (2002). Initial sequencing and comparative analysis of the mouse genome. Nature *420*, 520–562.

Mroczek, S., and Kufel, J. (2008). Apoptotic signals induce specific degradation of ribosomal RNA in yeast. Nucleic Acids Res. *36*, 2874–2888.

Mukaka, M. (2012). A guide to appropriate use of Correlation coefficient in medical research. Malawi Med. J. J. Med. Assoc. Malawi *24*, 69–71.

Naftelberg, S., Schor, I.E., Ast, G., and Kornblihtt, A.R. (2015). Regulation of Alternative Splicing Through Coupling with Transcription and Chromatin Structure. Annu. Rev. Biochem. *84*, 165–198.

Naito, K., Cho, E., Yang, G., Campbell, M.A., Yano, K., Okumoto, Y., Tanisaka, T., and Wessler, S.R. (2006). Dramatic amplification of a rice transposable element during recent domestication. Proc. Natl. Acad. Sci. *103*, 17620–17625.

Naito, K., Zhang, F., Tsukiyama, T., Saito, H., Hancock, C.N., Richardson, A.O., Okumoto, Y., Tanisaka, T., and Wessler, S.R. (2009). Unexpected consequences of a sudden and massive transposon amplification on rice gene expression. Nature *461*, 1130–1134.

Naumann, K., Fischer, A., Hofmann, I., Krauss, V., Phalke, S., Irmler, K., Hause, G., Aurich, A.-C., Dorn, R., Jenuwein, T., et al. (2005). Pivotal role of AtSUVH2 in heterochromatic histone methylation and gene silencing in Arabidopsis. EMBO J. *24*, 1418–1429.

Nekrutenko, A., and Li, W.-H. (2001). Transposable elements are found in a large number of human protein-coding genes. Trends Genet. *17*, 619–621.

Nesbitt, A., Kemp, B., Steele, C., Lovett, A., and Dorling, S. (2016). Impact of recent climate change and weather variability on the viability of UK viticulture – combining weather and climate records with producers' perspectives. Aust. J. Grape Wine Res. *22*, 324–335.

Neumann, P., Novák, P., Hoštáková, N., and Macas, J. (2019). Systematic survey of plant LTR-retrotransposons elucidates phylogenetic relationships of their polyprotein domains and provides a reference for element classification. Mob. DNA *10*, 1.

Nguyen, C.T., Tran, G.-B., and Nguyen, N.H. (2020). Homeostasis of histone acetylation is critical for auxin signaling and root morphogenesis. Plant Mol. Biol. *103*, 1–7.

Nigumann, P., Redik, K., Mätlik, K., and Speek, M. (2002). Many Human Genes Are Transcribed from the Antisense Promoter of L1 Retrotransposon. Genomics *79*, 628–634.

NobelPrize.org Barbara McClintock – Prize presentation. https://www.nobelprize.org/prizes/medicine/1983/mcclintock/prize-presentation/.

Nystedt, B., Street, N.R., Wetterbom, A., Zuccolo, A., Lin, Y.-C., Scofield, D.G., Vezzi, F., Delhomme, N., Giacomello, S., Alexeyenko, A., et al. (2013). The Norway spruce genome sequence and conifer genome evolution. Nature *497*, 579–584.

Oberlin, S., Sarazin, A., Chevalier, C., Voinnet, O., and Marí-Ordóñez, A. (2017). A genome-wide transcriptome and translatome analysis of *Arabidopsis* transposons identifies a unique and conserved genome expression strategy for *Ty1/Copia* retroelements. Genome Res. *27*, 1549–1562.

O'Donnell, K.A., and Burns, K.H. (2010). Mobilizing diversity: transposable element insertions in genetic variation and disease. Mob. DNA *1*, 21.

Okonechnikov, K., Golosova, O., and Fursov, M. (2012). Unipro UGENE: a unified bioinformatics toolkit. Bioinformatics *28*, 1166–1167.

Ong-Abdullah, M., Ordway, J.M., Jiang, N., Ooi, S.-E., Kok, S.-Y., Sarpan, N., Azimi, N., Hashim, A.T., Ishak, Z., Rosli, S.K., et al. (2015). Loss of Karma transposon methylation underlies the mantled somaclonal variant of oil palm. Nature *525*, 533–537.

Oxford Nanopore Technologies (2020a). Pychopper: A tool to identify, orient, trim and rescue full length cDNA reads (Oxford Nanopore Technologies).

Oxford Nanopore Technologies (2020b). Pinfish: tools to annotate genomes using long read transcriptomics data (Oxford Nanopore Technologies).

Panda, K., and Slotkin, R.K. (2020). Long-read cDNA Sequencing Enables a 'Gene-Like' Transcript Annotation of Arabidopsis Transposable Elements (Plant Biology).

Panda, K., Ji, L., Neumann, D.A., Daron, J., Schmitz, R.J., and Slotkin, R.K. (2016). Full-length autonomous transposable elements are preferentially targeted by expression-dependent forms of RNA-directed DNA methylation. Genome Biol. *17*, 170.

Pandey, R., Müller, A., Napoli, C.A., Selinger, D.A., Pikaard, C.S., Richards, E.J., Bender, J., Mount, D.W., and Jorgensen, R.A. (2002). Analysis of histone acetyltransferase and histone deacetylase families of Arabidopsis thaliana suggests functional diversification of chromatin modification among multicellular eukaryotes. Nucleic Acids Res. *30*, 5036–5055.

Parker, M.T., Knop, K., Sherwood, A.V., Schurch, N.J., Mackinnon, K., Gould, P.D., Hall, A.J., Barton, G.J., and Simpson, G.G. (2020). Nanopore direct RNA sequencing maps the complexity of Arabidopsis mRNA processing and m6A modification. ELife *9*, e49658.

Pelsy, F. (2010). Molecular and cellular mechanisms of diversity within grapevine varieties. Heredity *104*, 331–340.

Penterman, J., Zilberman, D., Huh, J.H., Ballinger, T., Henikoff, S., and Fischer, R.L. (2007). DNA demethylation in the Arabidopsis genome. Proc. Natl. Acad. Sci. *104*, 6752–6757.

Peschke, V.M., Phillips, R.L., and Gengenbach, B.G. (1987). Discovery of Transposable Element Activity Among Progeny of Tissue Culture—Derived Maize Plants. Science *238*, 804–807.

Pfluger, J., and Wagner, D. (2007). Histone modifications and dynamic regulation of genome accessibility in plants. Curr. Opin. Plant Biol. *10*, 645–652.

Pinedo, I., Ledger, T., Greve, M., and Poupin, M.J. (2015). Burkholderia phytofirmans PsJN induces long-term metabolic and transcriptional changes involved in Arabidopsis thaliana salt tolerance. Front. Plant Sci. *6*.

Plant and Food Research (2019). TEFingerprint, https://github.com/PlantandFoodResearch/TEFingerprint.

Pontier, D., Picart, C., Roudier, F., Garcia, D., Lahmy, S., Azevedo, J., Alart, E., Laudié, M., Karlowski, W.M., Cooke, R., et al. (2012). NERD, a plant-specific GW protein, defines an additional RNAi-dependent chromatin-based pathway in Arabidopsis. Mol. Cell *48*, 121–132.

Qi, F., and Zhang, F. (2020). Cell Cycle Regulation in the Plant Response to Stress. Front. Plant Sci. *10*.

Qi, Y., Denli, A.M., and Hannon, G.J. (2005). Biochemical Specialization within Arabidopsis RNA Silencing Pathways. Mol. Cell *19*, 421–428.

Quadrana, L., Almeida, J., Asís, R., Duffy, T., Dominguez, P.G., Bermúdez, L., Conti, G., Corrêa da Silva, J.V., Peralta, I.E., Colot, V., et al. (2014). Natural occurring epialleles determine vitamin E accumulation in tomato fruits. Nat. Commun. *5*, 4027.

Quadrana, L., Etcheverry, M., Gilly, A., Caillieux, E., Madoui, M.-A., Guy, J., Bortolini Silveira, A., Engelen, S., Baillet, V., Wincker, P., et al. (2019). Transposition favors the generation of large effect mutations that may facilitate rapid adaption. Nat. Commun. *10*, 3421.

Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841–842.

Rakocevic, A., Mondy, S., Tirichine, L., Cosson, V., Brocard, L., Iantcheva, A., Cayrel, A., Devier, B., Abu El-Heba, G.A., and Ratet, P. (2009). *MERE1* , a Low-Copy-Number Copia-Type Retroelement in *Medicago truncatula* Active during Tissue Culture. Plant Physiol. *151*, 1250–1263.

Rang, F.J., Kloosterman, W.P., and de Ridder, J. (2018). From squiggle to basepair: computational approaches for improving nanopore sequencing read accuracy. Genome Biol. *19*, 90.

Rangani, G., Underwood, J.L., and Srivastava, V. (2015). Chromatin analysis of an Arabidopsis Phytochrome A allele reveals the correlation of transcriptional repression with recalcitrance to histone acetylation. Plant Growth Regul. *75*, 179–186.

Ransom, R.F., and Walton, J.D. (1997). Histone Hyperacetylation in Maize in Response to Treatment with HC-Toxin or Infection by the Filamentous Fungus Cochliobolus carbonum. Plant Physiol. *115*, 1021–1027.

Ravindran, S. (2012). Barbara McClintock and the discovery of jumping genes. Proc. Natl. Acad. Sci. *109*, 20198–20199.

Rech, G.E., Bogaerts-Márquez, M., Barrón, M.G., Merenciano, M., Villanueva-Cañas, J.L., Horváth, V., Fiston-Lavier, A.-S., Luyten, I., Venkataram, S., Quesneville, H., et al. (2019). Stress response, behavior, and development are shaped by transposable element-induced mutations in Drosophila. PLOS Genet. *15*, e1007900.

Reddy, A.S.N., Rogers, M.F., Richardson, D.N., Hamilton, M., and Ben-Hur, A. (2012). Deciphering the Plant Splicing Code: Experimental and Computational Approaches for Predicting Alternative Splicing and Splicing Regulatory Elements. Front. Plant Sci. *3*.

Rey, O., Danchin, E., Mirouze, M., Loot, C., and Blanchet, S. (2016). Adaptation to Global Change: A Transposable Element–Epigenetics Perspective. Trends Ecol. Evol. *31*, 514–526.

Roberts, A., Trapnell, C., Donaghey, J., Rinn, J.L., and Pachter, L. (2011). Improving RNA-Seq expression estimates by correcting for fragment bias. Genome Biol. *12*, R22.

Roudier, F., Ahmed, I., Bérard, C., Sarazin, A., Mary-Huard, T., Cortijo, S., Bouyer, D., Caillieux, E., Duvernois-Berthet, E., Al-Shikhley, L., et al. (2011). Integrative epigenomic mapping defines four main chromatin states in Arabidopsis: Organization of the Arabidopsis epigenome. EMBO J. *30*, 1928–1938.

Ryan, D.P., and Owen-Hughes, T. (2011). Snf2-family proteins: chromatin remodellers for any occasion. Curr. Opin. Chem. Biol. *15*, 649–656.

Rymen, B., Kawamura, A., Lambolez, A., Inagaki, S., Takebayashi, A., Iwase, A., Sakamoto, Y., Sako, K., Favero, D.S., Ikeuchi, M., et al. (2019). Histone acetylation orchestrates wound-induced transcriptional activation and cellular reprogramming in Arabidopsis. Commun. Biol. *2*, 1–15.

Ryu, H., Cho, H., Bae, W., and Hwang, I. (2014). Control of early seedling development by BES1/TPL/HDA19-mediated epigenetic regulation of ABI3. Nat. Commun. *5*, 4138.

Saint-André, V., Batsché, E., Rachez, C., and Muchardt, C. (2011). Histone H3 lysine 9 trimethylation and HP1γ favor inclusion of alternative exons. Nat. Struct. Mol. Biol. *18*, 337–344.

Sakai, H., Tanaka, T., and Itoh, T. (2007). Birth and death of genes promoted by transposable elements in Oryza sativa. Gene *392*, 59–63.

SanMiguel, P., Tikhonov, A., Jin, Y.-K., Motchoulskaia, N., Zakharov, D., Melake-Berhan, A., Springer, P.S., Edwards, K.J., Lee, M., Avramova, Z., et al. (1996). Nested Retrotransposons in the Intergenic Regions of the Maize Genome. Science *274*, 765–768.

SanMiguel, P., Gaut, B.S., Tikhonov, A., Nakajima, Y., and Bennetzen, J.L. (1998). The paleontology of intergene retrotransposons of maize. Nat. Genet. *20*, 43–45.

Santos, J.A., Fraga, H., Malheiro, A.C., Moutinho-Pereira, J., Dinis, L.-T., Correia, C., Moriondo, M., Leolini, L., Dibari, C., Costafreda-Aumedes, S., et al. (2020). A Review of the Potential Climate Change Impacts and Adaptation Options for European Viticulture. Appl. Sci. *10*, 3092.

Saze, H. (2018). Epigenetic regulation of intragenic transposable elements: a two-edged sword. J. Biochem. (Tokyo) *164*, 323–328.

Saze, H., Shiraishi, A., Miura, A., and Kakutani, T. (2008). Control of Genic DNA Methylation by a jmjC Domain–Containing Protein in. *319*, 5.

Saze, H., Kitayama, J., Takashima, K., Miura, S., Harukawa, Y., Ito, T., and Kakutani, T. (2013). Mechanism for full-length RNA processing of Arabidopsis genes containing intragenic heterochromatin. Nat. Commun. *4*, 2301.

Schnable, P.S., Ware, D., Fulton, R.S., Stein, J.C., Wei, F., Pasternak, S., Liang, C., Zhang, J., Fulton, L., Graves, T.A., et al. (2009). The B73 maize genome: complexity, diversity, and dynamics. Science *326*, 1112–1115.

Schorn, A.J., and Martienssen, R. (2018). Tie-Break: Host and Retrotransposons Play tRNA. Trends Cell Biol. *28*, 793–806.

Schorn, A.J., Gutbrod, M.J., LeBlanc, C., and Martienssen, R. (2017). LTR-Retrotransposon Control by tRNA-Derived Small RNAs. Cell *170*, 61-71.e11.

Schulman, A.H. (2013). Retrotransposon replication in plants. Curr. Opin. Virol. *3*, 604–614.

Secco, D., Wang, C., Shou, H., Schultz, M.D., Chiarenza, S., Nussaume, L., Ecker, J.R., Whelan, J., and Lister, R. (2015). Stress induced gene expression drives transient DNA methylation changes at adjacent repetitive elements. ELife *4*, e09343.

Sedlazeck, F.J., Lee, H., Darby, C.A., and Schatz, M.C. (2018). Piercing the dark matter: bioinformatics of long-range sequencing and mapping. Nat. Rev. Genet. *19*, 329–346.

Sessegolo, C., Cruaud, C., Da Silva, C., Cologne, A., Dubarry, M., Derrien, T., Lacroix, V., and Aury, J.-M. (2019). Transcriptome profiling of mouse samples using nanopore sequencing of cDNA and RNA molecules. Sci. Rep. *9*, 14908.

Settles, A.M., Baron, A., Barkan, A., and Martienssen, R.A. (2001). Duplication and Suppression of Chloroplast Protein Translocation Genes in Maize. Genetics *157*, 349–360.

Shahbazian, M.D., and Grunstein, M. (2007). Functions of Site-Specific Histone Acetylation and Deacetylation. Annu. Rev. Biochem. *76*, 75–100.

Shahid, S., and Slotkin, R.K. (2020). The current revolution in transposable element biology enabled by long reads. Curr. Opin. Plant Biol. *54*, 49–56.

Sharma, U., Conine, C.C., Shea, J.M., Boskovic, A., Derr, A.G., Bing, X.Y., Belleannee, C., Kucukural, A., Serra, R.W., Sun, F., et al. (2016). Biogenesis and function of tRNA fragments during sperm maturation and fertilization in mammals. Science *351*, 391–396.

Sigman, M.J., and Slotkin, R.K. (2016). The First Rule of Plant Transposable Element Silencing: Location, Location, Location. Plant Cell *28*, 304–313.

Singer, T. (2001). Robertson's Mutator transposons in A. thaliana are regulated by the chromatin-remodeling gene Decrease in DNA Methylation (DDM1). Genes Dev. *15*, 591–602.

Slotkin, R.K., and Martienssen, R. (2007). Transposable elements and the epigenetic regulation of the genome. Nat. Rev. Genet. *8*, 272–285.

Slotkin, R.K., Vaughn, M., Borges, F., Tanurdžić, M., Becker, J.D., Feijó, J.A., and Martienssen, R.A. (2009). Epigenetic Reprogramming and Small RNA Silencing of Transposable Elements in Pollen. Cell *136*, 461–472.

Soneson, C., and Delorenzi, M. (2013). A comparison of methods for differential expression analysis of RNA-seq data. BMC Bioinformatics *14*, 91.

Springer, N.M., and Schmitz, R.J. (2017). Exploiting induced and natural epigenetic variation for crop improvement. Nat. Rev. Genet. *18*, 563–575.

Stasevich, T.J., Hayashi-Takanaka, Y., Sato, Y., Maehara, K., Ohkawa, Y., Sakata-Sogawa, K., Tokunaga, M., Nagase, T., Nozaki, N., McNally, J.G., et al. (2014). Regulation of RNA polymerase II activation by histone acetylation in single living cells. Nature *516*, 272–275.

Stroud, H., Do, T., Du, J., Zhong, X., Feng, S., Johnson, L., Patel, D.J., and Jacobsen, S.E. (2014). Non-CG methylation patterns shape the epigenetic landscape in Arabidopsis. Nat. Struct. Mol. Biol. *21*, 64–72.

Studer, A., Zhao, Q., Ross-Ibarra, J., and Doebley, J. (2011). Identification of a functional transposon insertion in the maize domestication gene tb1. Nat. Genet. *43*, 1160–1163.

Sturman, A., and Quénol, H. (2012). Changes in atmospheric circulation and temperature trends in major vineyard regions of New Zealand: ATMOSPHERIC CIRCULATION AND TEMPERATURE IN NEW ZEALAND VINEYARD AREAS. Int. J. Climatol. n/a-n/a.

Takahashi, S., and Murata, N. (2008). How do environmental stresses accelerate photoinhibition? Trends Plant Sci. *13*, 178–182.

Takeda, S., Sugimoto, K., Otsuki, H., and Hirochika, H. (1999). A 13-bp cis-regulatory element in the LTR promoter of the tobacco retrotransposon Tto1 is involved in responsiveness to tissue culture, wounding, methyl jasmonate and fungal elicitors. Plant J. *18*, 383–393.

Tang, A.D., Soulette, C.M., van Baren, M.J., Hart, K., Hrabeta-Robinson, E., Wu, C.J., and Brooks, A.N. (2020). Full-length transcript characterization of SF3B1 mutation in chronic lymphocytic leukemia reveals downregulation of retained introns. Nat. Commun. *11*, 1438.

Tanurdzic, M., Vaughn, M.W., Jiang, H., Lee, T.-J., Slotkin, R.K., Sosinski, B., Thompson, W.F., Doerge, R.W., and Martienssen, R.A. (2008). Epigenomic Consequences of Immortalized Plant Cell Suspension Culture. PLOS Biol. *6*, e302.

Tenaillon, M.I., Hollister, J.D., and Gaut, B.S. (2010). A triptych of the evolution of plant transposable elements. Trends Plant Sci. *15*, 471–478.

Tenhaken, R. (2015). Cell wall remodeling under abiotic stress. Front. Plant Sci. *5*.

Tessadori, F., Zanten, M. van, Pavlova, P., Clifton, R., Pontvianne, F., Snoek, L.B., Millenaar, F.F., Schulkes, R.K., Driel, R. van, Voesenek, L.A.C.J., et al. (2009). PHYTOCHROME B and HISTONE DEACETYLASE 6 Control Light-Induced Chromatin Compaction in Arabidopsis thaliana. PLOS Genet. *5*, e1000638.

Thiagalingam, S., Cheng, K.-H., Lee, H.J., Mineva, N., Thiagalingam, A., and Ponte, J.F. (2003). Histone Deacetylases: Unique Players in Shaping the Epigenetic Histone Code. Ann. N. Y. Acad. Sci. *983*, 84–100.

Thieme, M., and Bucher, E. (2018). Transposable Elements as Tool for Crop Improvement. In Advances in Botanical Research, (Elsevier), pp. 165–202.

Thieme, M., Lanciano, S., Balzergue, S., Daccord, N., Mirouze, M., and Bucher, E. (2017). Inhibition of RNA polymerase II allows controlled mobilisation of retrotransposons for plant breeding. Genome Biol. *18*, 134.

Thompson, D.M., Lu, C., Green, P.J., and Parker, R. (2008). tRNA cleavage is a conserved response to oxidative stress in eukaryotes. RNA *14*, 2095–2103.

Tiricz, H., Nagy, B., Ferenc, G., Török, K., Nagy, I., Dudits, D., and Ayaydin, F. (2018). Relaxed chromatin induced by histone deacetylase inhibitors improves the oligonucleotide-directed gene editing in plant cells. J. Plant Res. *131*, 179–189.

Tittel-Elmer, M., Bucher, E., Broger, L., Mathieu, O., Paszkowski, J., and Vaillant, I. (2010). Stress-Induced Activation of Heterochromatic Transcription. PLOS Genet. *6*, e1001175.

To, T.K., Kim, J.-M., Matsui, A., Kurihara, Y., Morosawa, T., Ishida, J., Tanaka, M., Endo, T., Kakutani, T., Toyoda, T., et al. (2011a). Arabidopsis HDA6 Regulates Locus-Directed Heterochromatin Silencing in Cooperation with MET1. PLoS Genet. *7*, e1002055.

To, T.K., Kim, J.-M., Matsui, A., Kurihara, Y., Morosawa, T., Ishida, J., Tanaka, M., Endo, T., Kakutani, T., Toyoda, T., et al. (2011b). Arabidopsis HDA6 Regulates Locus-Directed Heterochromatin Silencing in Cooperation with MET1. PLOS Genet. *7*, e1002055.

Torregrosa, L. (1998). A simple and efficient method to obtain stable embryogenic cultures from anthers of Vitis vinifera L. Vitis *37*, 91–92.

Trapnell, C., Williams, B.A., Pertea, G., Mortazavi, A., Kwan, G., Baren, M.J. van, Salzberg, S.L., Wold, B.J., and Pachter, L. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat. Biotechnol. *28*, 511–515.

Tsuchiya, T., and Eulgem, T. (2013). An alternative polyadenylation mechanism coopted to the Arabidopsis RPP7 gene through intronic retrotransposon domestication. Proc. Natl. Acad. Sci. *110*, E3535–E3543.

Tsukahara, S., Kobayashi, A., Kawabe, A., Mathieu, O., Miura, A., and Kakutani, T. (2009). Bursts of retrotransposition reproduced in Arabidopsis. Nature *461*, 423–426.

Turner, B.M. (2000). Histone acetylation and an epigenetic code. BioEssays *22*, 836–845.

Valencia-Sanchez, M.A., Liu, J., Hannon, G.J., and Parker, R. (2006). Control of translation and mRNA degradation by miRNAs and siRNAs. Genes Dev. *20*, 515–524.

Van't Hof, A.E., Campagne, P., Rigden, D.J., Yung, C.J., Lingley, J., Quail, M.A., Hall, N., Darby, A.C., and Saccheri, I.J. (2016). The industrial melanism mutation in British peppered moths is a transposable element. Nature *534*, 102–105.

Varagona, M.J., Purugganan, M., and Wessler, S.R. (1992). Alternative splicing induced by insertion of retrotransposons into the maize waxy gene. Plant Cell *4*, 811–820.

Vaser, R., Sović, I., Nagarajan, N., and Šikić, M. (2017). Fast and accurate de novo genome assembly from long uncorrected reads. Genome Res. *27*, 737–746.

Venables, W.N., and Ripley, B.D. (2002). Modern Applied Statistics with S (New York: Springer).

Vicient, C.M. (2010). Transcriptional activity of transposable elements in maize. BMC Genomics *11*, 601.

Vitte, C., Chaparro, C., Quesneville, H., and Panaud, O. (2007). Spip and Squiq, two novel rice non-autonomous LTR retro-element families related to RIRE3 and RIRE8. Plant Sci. *172*, 8–19.

Volden, R., Palmer, T., Byrne, A., Cole, C., Schmitz, R.J., Green, R.E., and Vollmers, C. (2018). Improving nanopore read accuracy with the R2C2 method enables the sequencing of highly multiplexed full-length single-cell cDNA. Proc. Natl. Acad. Sci. *115*, 9726–9731.

Wang, I.-F., Wu, L.-S., and Shen, C.-K.J. (2008). TDP-43: an emerging new player in neurodegenerative diseases. Trends Mol. Med. *14*, 479–485.

Wang, X., Duan, C.-G., Tang, K., Wang, B., Zhang, H., Lei, M., Lu, K., Mangrauthia, S.K., Wang, P., Zhu, G., et al. (2013). RNA-binding protein regulates plant DNA methylation by controlling mRNA processing at the intronic heterochromatin-containing gene IBM1. Proc. Natl. Acad. Sci. *110*, 15467–15472.

Wang, Z., Zang, C., Cui, K., Schones, D.E., Barski, A., Peng, W., and Zhao, K. (2009). Genome-wide Mapping of HATs and HDACs Reveals Distinct Functions in Active and Inactive Genes. Cell *138*, 1019–1031.

Warnes, G.R., Bolker, B., Bonebakker, L., Gentleman, R., Huber, W., Liaw, A., Lumley, T., Maechler, M., Magnusson, A., Moeller, S., et al. (2020). gplots: Various R Programming Tools for Plotting Data.

Watanabe, N., and Lam, E. (2008). BAX Inhibitor-1 Modulates Endoplasmic Reticulum Stress-mediated Programmed Cell Death in Arabidopsis. J. Biol. Chem. *283*, 3200–3210.

Waterborg, J.H., and Kapros, T. (2002). Kinetic analysis of histone acetylation turnover and Trichostatin A induced hyper- and hypoacetylation in alfalfa. Biochem. Cell Biol. *80*, 279–293.

Weirather, J.L., de Cesare, M., Wang, Y., Piazza, P., Sebastiano, V., Wang, X.-J., Buck, D., and Au, K.F. (2017). Comprehensive comparison of Pacific Biosciences and Oxford Nanopore Technologies and their applications to transcriptome analysis. F1000Research *6*.

Wenger, A.M., Peluso, P., Rowell, W.J., Chang, P.-C., Hall, R.J., Concepcion, G.T., Ebler, J., Fungtammasan, A., Kolesnikov, A., Olson, N.D., et al. (2019). Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. Nat. Biotechnol. *37*, 1155–1162.

Wessler, S.R. (2006). Transposable elements and the evolution of eukaryotic genomes. Proc. Natl. Acad. Sci. *103*, 17600–17601.

West, P.T., Li, Q., Ji, L., Eichten, S.R., Song, J., Vaughn, M.W., Schmitz, R.J., and Springer, N.M. (2014). Genomic Distribution of H3K9me2 and DNA Methylation in a Maize Genome. PLoS ONE *9*, e105267.

Wick, R.R., Judd, L.M., and Holt, K.E. (2019). Performance of neural network basecalling tools for Oxford Nanopore sequencing. Genome Biol. *20*, 129.

Wicker, T., Sabot, F., Hua-Van, A., Bennetzen, J.L., Capy, P., Chalhoub, B., Flavell, A., Leroy, P., Morgante, M., Panaud, O., et al. (2007). A unified classification system for eukaryotic transposable elements. Nat. Rev. Genet. *8*, 973–982.

Wickham, H. (2016). ggplot2: Elegant Graphics for Data Analysis (Springer-Verlag New York).

Wickham, H., François, R., Henry, L., and Müller, K. (2018). dplyr: A Grammar of Data Manipulation, https://CRAN.R-project.org/package=dplyr.

Wierzbicki, A.T., Haag, J.R., and Pikaard, C.S. (2008). Noncoding transcription by RNA Polymerase Pol IVb/Pol V mediates transcriptional silencing of overlapping and adjacent genes. Cell *135*, 635–648.

Wiley, J.C., Meabon, J.S., Frankowski, H., Smith, E.A., Schecterson, L.C., Bothwell, M., and Ladiges, W.C. (2010). Phenylbutyric Acid Rescues Endoplasmic Reticulum Stress-Induced Suppression of APP Proteolysis and Prevents Apoptosis in Neuronal Cells. PLOS ONE *5*, e9135.

Witt, O., Deubzer, H.E., Milde, T., and Oehme, I. (2009). HDAC family: What are the cancer relevant targets? Cancer Lett. *277*, 8–21.

Workman, R.E., Tang, A.D., Tang, P.S., Jain, M., Tyson, J.R., Razaghi, R., Zuzarte, P.C., Gilpatrick, T., Payne, A., Quick, J., et al. (2019). Nanopore native RNA sequencing of a human poly(A) transcriptome. Nat. Methods *16*, 1297–1305.

Wu, T.D., and Watanabe, C.K. (2005). GMAP: a genomic mapping and alignment program for mRNA and EST sequences. Bioinformatics *21*, 1859–1875.

Wu, L., Zhou, H., Zhang, Q., Zhang, J., Ni, F., Liu, C., and Qi, Y. (2010). DNA Methylation Mediated by a MicroRNA Pathway. Mol. Cell *38*, 465–475.

Yamashita, H., and Tahara, M. (2006). A LINE-type Retrotransposon Active in Meristem Stem Cells Causes Heritable Transpositions in the Sweet Potato Genome. Plant Mol. Biol. *61*, 79–84.

Yang, N., and Kazazian, H.H. (2006). L1 retrotransposition is suppressed by endogenously encoded small interfering RNAs in human cultured cells. Nat. Struct. Mol. Biol. *13*, 763–771.

Yang, G., Lee, Y.-H., Jiang, Y., Shi, X., Kertbundit, S., and Hall, T.C. (2005). A Two-Edged Role for the Transposable Element *Kiddo* in the *rice ubiquitin2* Promoter. Plant Cell *17*, 1559–1568.

Yang, X., Srivastava, R., Howell, S.H., and Bassham, D.C. (2016). Activation of autophagy by unfolded proteins during endoplasmic reticulum stress. Plant J. *85*, 83–95.

Ye, R., Chen, Z., Lian, B., Rowley, M.J., Xia, N., Chai, J., Li, Y., He, X.-J., Wierzbicki, A.T., and Qi, Y. (2016). A Dicer-Independent Route for Biogenesis of siRNAs that Direct DNA Methylation in Arabidopsis. Mol. Cell *61*, 222–235.

Yu, A., Lepere, G., Jay, F., Wang, J., Bapaume, L., Wang, Y., Abraham, A.-L., Penterman, J., Fischer, R.L., Voinnet, O., et al. (2013). Dynamics and biological relevance of DNA demethylation in Arabidopsis antibacterial defense. Proc. Natl. Acad. Sci. *110*, 2389–2394.

Yu, C.-W., Liu, X., Luo, M., Chen, C., Lin, X., Tian, G., Lu, Q., Cui, Y., and Wu, K. (2011). HISTONE DEACETYLASE6 Interacts with FLOWERING LOCUS D and Regulates Flowering in Arabidopsis. Plant Physiol. *156*, 173–184.

Yu, C.-W., Tai, R., Wang, S.-C., Yang, P., Luo, M., Yang, S., Cheng, K., Wang, W.-C., Cheng, Y.-S., and Wu, K. (2017). HISTONE DEACETYLASE6 Acts in Concert with Histone Methyltransferases SUVH4, SUVH5, and SUVH6 to Regulate Transposon Silencing. Plant Cell *29*, 1970–1983.

Yu, X., Li, L., Li, L., Guo, M., Chory, J., and Yin, Y. (2008). Modulation of brassinosteroid-regulated gene expression by jumonji domain-containing proteins ELF6 and REF6 in Arabidopsis. Proc. Natl. Acad. Sci. *105*, 7618–7623.

Zemach, A., Kim, M.Y., Hsieh, P.-H., Coleman-Derr, D., Eshed-Williams, L., Thao, K., Harmer, S.L., and Zilberman, D. (2013). The Arabidopsis Nucleosome Remodeler DDM1 Allows DNA Methyltransferases to Access H1-Containing Heterochromatin. Cell *153*, 193–205.

Zentner, G.E., and Henikoff, S. (2013). Regulation of nucleosome dynamics by histone modifications. Nat. Struct. Mol. Biol. *20*, 259–266.

Zhang, H., Lang, Z., and Zhu, J.-K. (2018). Dynamics and function of DNA methylation in plants. Nat. Rev. Mol. Cell Biol. *19*, 489–506.

Zhang, X., Zhao, H., Gao, S., Wang, W.-C., Katiyar-Agarwal, S., Huang, H.-D., Raikhel, N., and Jin, H. (2011). Arabidopsis Argonaute 2 Regulates Innate Immunity via miRNA393∗-Mediated Silencing of a Golgi-Localized SNARE Gene, MEMB12. Mol. Cell *42*, 356–366.

Zhao, D., Ferguson, A.A., and Jiang, N. (2016). What makes up plant genomes: The vanishing line between transposable elements and genes. Biochim. Biophys. Acta BBA - Gene Regul. Mech. *1859*, 366–380.

Zhou, C., Zhang, L., Duan, J., Miki, B., and Wu, K. (2005). HISTONE DEACETYLASE19 Is Involved in Jasmonic Acid and Ethylene Signaling of Pathogen Response in Arabidopsis. Plant Cell *17*, 1196–1204.

Zhu, J., Kapoor, A., Sridhar, V.V., Agius, F., and Zhu, J.-K. (2007). The DNA Glycosylase/Lyase ROS1 Functions in Pruning DNA Methylation Patterns in Arabidopsis. Curr. Biol. *17*, 54–59.

# Appendix A

# Supplementary remarks

## A.1 The workflow, algorithm, and the restriction of TEtranscirpts

The software TEtranscripts (Jin et al. 2015) estimates read abundance for each TE locus as follows:

(1) Calculate 'initial read abundance':

Firstly TEtranscripts calculates the 'initial read abundance' which applies the equal-weighting principle to assign multi-mapping reads evenly to the associated TE loci.

(2) Calculate 'adjusted read abundance':

For each TE locus, TEtranscripts calculates the 'adjusted read abundance' by normalizing the equal-weighting abundances (i.e. 'initial read abundance') with TE's 'effective length', which is determined by the annotated length of the TE locus and the fragment size of sequencing library (Trapnell et al., 2010).

The idea of 'effective length' was first brought up by Trapnell et al. (2010) to work out the bias introduced from the size of a mapped feature. By presenting more mappable locations, the TE locus that is longer in length may be assigned with more equal-weighted multi-mapping reads. To take the size of the TE locus into consideration, normalizing the 'initial read abundance' by 'effective length' may reduce the bias in read quantification caused by the length of the feature.

The 'effective length' of a TE locus $t$ is defined as $\tilde{l}_t = l_t - m + 1$, where $l_t$ denotes the length of the TE locus, and $m$ denotes the sequenced library's fragment length, which can be inferred from the distance of alignment positions between paired reads (Jin et al., 2015). In other words, the effective length $\tilde{l}_t$ of the TE locus $t$ mathematically represents the number of possible start sites the TE locus could have generated a sequencing fragment of that particular length $m$ (Harold Pimentel, 2014; Trapnell et al., 2010). Imaging that a TE locus spanning 1 kb in the genome is mapped with a pair of reads that are sequenced from a fragment of 500 bp in length; disregarding the sequence context, there're 501 possible sites (calculated as $1000 - 500 + 1 = 501$) on this TE locus to produce the 500-bp sequencing fragment. Hence, this TE's 'initial read abundance', which is the sum of uniformly divided multi-mapping reads, is further divided by the 'effective length' to generate an 'adjusted read

abundance' that is normalized firstly by the number of mapping sites and secondly by the 'effective length' of the TE locus. The 'initial relative abundance' is then computed using the 'adjusted read abundance'.

(3) Calculate 'initial relative abundance':

TEtranscripts then calculates the 'initial relative abundance', where the 'adjusted read abundance' (obtained from the previous step) of each TE locus is divided by the sum of the 'adjusted read abundances' from all TE loci. The 'initial relative abundance' is then utilized in the first iteration cycle of the following algorithm.

(4) Perform expectation-maximization (EM) algorithm:

The EM algorithm of TEtranscripts alternatively runs the expectation step (E-step) that re-estimate 'adjusted read abundance', and the maximization step (M-step) that re-optimises the 'relative abundance' of each TE locus until the estimated 'relative abundance' converge (Jin et al. 2015).

After multiple iterations of the EM algorithm, TE loci having shorter 'effective length' or more 'initial read abundance' would tend to have more 'relative abundance' than other TE loci that share multi-mapping reads with them. This might serve as a potential bias on closely related TE loci but can be minimized by the next step.

(5) Sum the 'relative abundance' at the TE family level:

The final 'relative abundance' of closely related TE loci (i.e. TE loci of the same family) are summed up as the total counts contributed from the corresponding TE family. However, the individual origins (TE loci) that generate the TE transcripts remain unknown.

# Appendix B

# Recipes for plant tissue culture media

## B.1 Recipes for plat tissue culture media

**Table B.1 Ingredients for plant tissue culture media**

| Medium | Reagent | Final conc. | Amount | Note | Reference |
|---|---|---|---|---|---|
| $C_1^P$ Medium (1 litre) | MS basal salt mix | 0.5 X | 2.2 g | pH6.0, autoclave | (Torregrosa, 1998) |
| | MS micro elements | 0.5 X | 0.5 g | | |
| | Casein hydrolysate | 0.1 % (w/v) | 1 g | | |
| | Sucrose | 3 % (w/v) | 30 g | | |
| | Gelrite | 0.5 % (w/v) | 5 g | | |
| | T Vitamins 1000X stock | 1 X | 1 mL | | |
| | BAP (100mM) | 1 µM | 10 µL | | |
| | 2,4-D (100mM) | 5 µM | 50 µL | | |
| T Vitamins 1000X stock (5 mL) | Myo-inositol | | 250 mg | | (Torregrosa, 1998) |
| | Nicotinic acid | | 5 mg | | |
| | Pyridoxine-HCl | | 5 mg | | |
| | Thiamine-HCl | | 5 mg | | |
| | Ca-Pantothenoate | | 5 mg | | |
| | D-Biotin (1 mg/mL) | | 50 µL | | |

# Appendix C

# Supplementary Data

## C.1 Alignments of reads unmapped to the grapevine reference genome to *S. cerevisiae* and *H. uvarum*

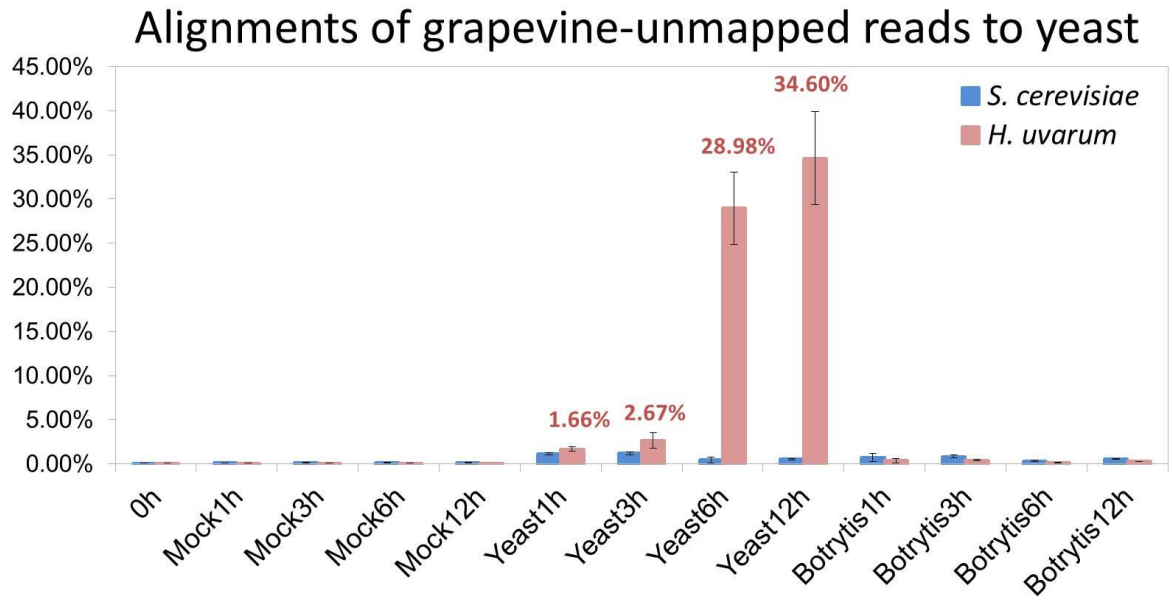



**Figure C. 1 Proportion of grapevine-unmapped reads aligned to *S. cerevisiae* and *H.uvarum***

Sequencing reads unmapped to the grapevine reference genome were mapped against the genomes of S. cerevisiae and H. uvarum. The proportion of reads mapped to yeast (y-axis) was calculated as the number of reads mapped to the indicated yeast species divided by the abundances of grapevine-unmapped reads. Each time-point was illustrated by the mean and the standard deviation (denoted by error bars) obtained from three technical replicates.

## C.2 Comparison of TE loci identity of four sets of expression candidate pools

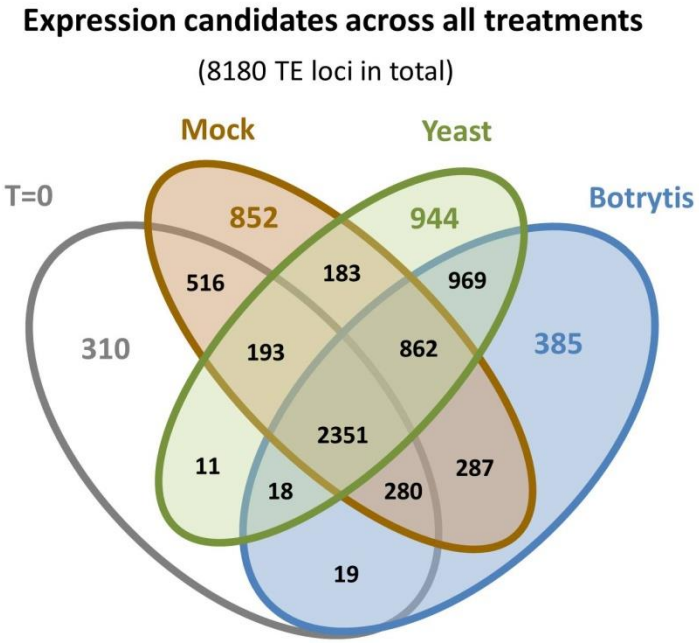



**Figure C. 2 Comparison of TE loci presented in four sets of expression candidates**

Different sets of expression candidates identified in T=0, mock, yeast and Botrytis treatments were, respectively, presented in grey, brown, green and blue ellipses. The overlapping areas include expression candidates that were presented in more than one of the four experimental conditions. For instance, the centre of this Venn diagram shows 2,351 expression candidates that were presented in all four experimental conditions.

## C.3 Expression candidates grouped by families, distinctness and integrity

### Table C.1 Copia expression candidates grouped by families, distinctness and integrity

FL and FG denote full-length and fragmented expression candidates, respectively.

| Copia Family | T=0 | | | | | Mock | | | | | Yeast | | | | | Botrytis | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Expression candidates | | | | | Expression candidates | | | | | Expression candidates | | | | | Expression candidates | | | | |
| | Untrackable | | Trackable | | Sum | Untrackable | | Trackable | | Sum | Untrackable | | Trackable | | Sum | Untrackable | | Trackable | | Sum |
| | FL | FG | FL | FG | | FL | FG | FL | FG | | FL | FG | FL | FG | | FL | FG | FL | FG | |
| Copia-10 | 1 | 2 | 2 | 17 | 22 | 0 | 1 | 3 | 17 | 21 | 0 | 0 | 1 | 23 | 24 | 0 | 0 | 1 | 25 | 26 |
| Copia-11 | 0 | 1 | 0 | 5 | 6 | 1 | 2 | 1 | 7 | 11 | 1 | 6 | 1 | 7 | 15 | 0 | 2 | 0 | 5 | 7 |
| Copia-12 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 |
| Copia-13 | 0 | 0 | 0 | 2 | 2 | 0 | 0 | 0 | 2 | 2 | 0 | 0 | 0 | 2 | 2 | 0 | 0 | 0 | 2 | 2 |
| Copia-15 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 1 | 1 |
| Copia-16 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 1 | 1 |
| Copia-17 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 2 | 2 |
| Copia-18A | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 4 | 4 | 0 | 0 | 0 | 10 | 10 | 0 | 0 | 0 | 5 | 5 |
| Copia-18 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 2 | 2 | 4 | 0 | 0 | 2 | 2 | 4 | 0 | 0 | 2 | 1 | 3 |
| Copia-19 | 0 | 0 | 0 | 2 | 2 | 0 | 0 | 0 | 3 | 3 | 0 | 2 | 0 | 2 | 4 | 0 | 2 | 0 | 1 | 3 |
| Copia-1A | 0 | 0 | 2 | 9 | 11 | 0 | 0 | 2 | 10 | 12 | 0 | 0 | 3 | 14 | 17 | 0 | 0 | 3 | 12 | 15 |
| Copia-1 | 0 | 1 | 2 | 20 | 23 | 0 | 0 | 4 | 22 | 26 | 0 | 2 | 5 | 25 | 32 | 0 | 1 | 5 | 28 | 34 |
| Copia-20 | 0 | 0 | 0 | 7 | 7 | 0 | 3 | 1 | 7 | 11 | 0 | 1 | 1 | 7 | 9 | 0 | 0 | 0 | 5 | 5 |
| Copia-21 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 |
| Copia-22 | 0 | 0 | 0 | 2 | 2 | 1 | 24 | 1 | 4 | 30 | 0 | 1 | 0 | 5 | 6 | 0 | 6 | 1 | 3 | 10 |
| Copia-23 | 118 | 75 | 14 | 4 | 211 | 136 | 115 | 26 | 12 | 289 | 126 | 115 | 41 | 14 | 296 | 126 | 93 | 26 | 11 | 256 |
| Copia-24 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 |
| Copia-26 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 1 | 2 | 3 | 0 | 0 | 1 | 2 | 3 |
| Copia-27 | 0 | 1 | 0 | 4 | 5 | 0 | 1 | 0 | 3 | 4 | 0 | 2 | 1 | 10 | 13 | 0 | 1 | 0 | 11 | 12 |
| Copia-28 | 0 | 6 | 2 | 5 | 13 | 0 | 8 | 2 | 6 | 16 | 0 | 7 | 2 | 5 | 14 | 0 | 6 | 2 | 6 | 14 |
| Copia-29b | 0 | 1 | 0 | 22 | 23 | 0 | 0 | 0 | 38 | 38 | 0 | 0 | 0 | 46 | 46 | 0 | 2 | 0 | 44 | 46 |
| Copia-29 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 1 | 1 |
| Copia-2 | 0 | 0 | 0 | 4 | 4 | 0 | 0 | 0 | 6 | 6 | 0 | 0 | 0 | 7 | 7 | 0 | 0 | 0 | 7 | 7 |
| Copia-30 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Copia-31 | 0 | 0 | 1 | 4 | 5 | 3 | 11 | 2 | 7 | 23 | 1 | 7 | 3 | 7 | 18 | 0 | 6 | 2 | 7 | 15 |
| Copia-32 | 0 | 2 | 0 | 2 | 4 | 0 | 2 | 0 | 2 | 4 | 5 | 33 | 0 | 1 | 39 | 0 | 9 | 0 | 2 | 11 |
| Copia-33 | 0 | 0 | 1 | 4 | 5 | 0 | 2 | 2 | 7 | 11 | 0 | 0 | 3 | 11 | 14 | 0 | 3 | 2 | 7 | 12 |
| Copia-34 | 1 | 1 | 0 | 6 | 8 | 5 | 8 | 1 | 7 | 21 | 4 | 9 | 1 | 7 | 21 | 6 | 5 | 0 | 6 | 17 |
| Copia-35 | 0 | 0 | 0 | 4 | 4 | 0 | 0 | 0 | 7 | 7 | 0 | 0 | 1 | 5 | 6 | 0 | 0 | 0 | 7 | 7 |
| Copia-36 | 0 | 2 | 0 | 3 | 5 | 0 | 4 | 0 | 4 | 8 | 0 | 0 | 0 | 2 | 2 | 0 | 0 | 0 | 3 | 3 |
| Copia-37 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 |
| Copia-38 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 2 | 0 | 1 | 0 | 3 | 4 | 0 | 0 | 0 | 1 | 1 |
| Copia-39 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Copia-3 | 53 | 30 | 2 | 4 | 89 | 61 | 69 | 14 | 9 | 153 | 57 | 68 | 15 | 14 | 154 | 59 | 70 | 11 | 10 | 150 |
| Copia-40 | 0 | 0 | 0 | 5 | 5 | 0 | 0 | 0 | 3 | 3 | 0 | 0 | 0 | 4 | 4 | 0 | 0 | 0 | 2 | 2 |
| Copia-41 | 0 | 0 | 1 | 2 | 3 | 0 | 0 | 2 | 3 | 5 | 0 | 0 | 3 | 3 | 6 | 0 | 0 | 2 | 3 | 5 |
| Copia-42 | 0 | 0 | 0 | 9 | 9 | 0 | 1 | 0 | 11 | 12 | 0 | 0 | 0 | 11 | 11 | 0 | 0 | 0 | 14 | 14 |
| Copia-43 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Copia-44 | 0 | 0 | 0 | 2 | 2 | 4 | 21 | 0 | 3 | 28 | 5 | 28 | 0 | 3 | 36 | 5 | 26 | 0 | 3 | 34 |
| Copia-45 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 2 | 0 | 0 | 0 | 3 | 3 | 0 | 0 | 0 | 0 | 0 |
| Copia-46 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 |
| Copia-47 | 0 | 28 | 0 | 4 | 32 | 0 | 34 | 1 | 9 | 44 | 0 | 5 | 0 | 2 | 7 | 0 | 26 | 1 | 2 | 29 |
| Copia-48 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Copia-49 | 0 | 10 | 0 | 7 | 17 | 0 | 26 | 0 | 12 | 38 | 0 | 28 | 0 | 11 | 39 | 0 | 33 | 0 | 13 | 46 |
| Copia-4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 |
| Copia-50 | 0 | 2 | 0 | 3 | 5 | 0 | 3 | 0 | 2 | 5 | 0 | 3 | 0 | 3 | 6 | 0 | 2 | 0 | 3 | 5 |
| Copia-51 | 0 | 2 | 0 | 4 | 6 | 0 | 1 | 0 | 4 | 5 | 0 | 0 | 0 | 4 | 4 | 0 | 1 | 0 | 3 | 4 |
| Copia-52 | 0 | 0 | 0 | 2 | 2 | 0 | 0 | 0 | 2 | 2 | 0 | 0 | 0 | 2 | 2 | 0 | 0 | 0 | 2 | 2 |
| Copia-53 | 0 | 0 | 0 | 6 | 6 | 0 | 1 | 0 | 8 | 9 | 0 | 0 | 0 | 12 | 12 | 0 | 0 | 0 | 6 | 6 |
| Copia-54 | 0 | 0 | 0 | 2 | 2 | 0 | 0 | 0 | 2 | 2 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 1 | 1 |
| Copia-55 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| Copia-56 | 0 | 0 | 1 | 14 | 15 | 0 | 0 | 1 | 17 | 18 | 0 | 0 | 1 | 24 | 25 | 0 | 0 | 1 | 25 | 26 |
| Copia-57 | 0 | 0 | 0 | 2 | 2 | 0 | 0 | 0 | 2 | 2 | 0 | 0 | 0 | 5 | 5 | 0 | 0 | 1 | 4 | 5 |
| Copia-58 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 | 3 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 |
| Copia-59 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 2 | 0 | 0 | 0 | 2 | 2 |
| Copia-5 | 0 | 0 | 0 | 2 | 2 | 0 | 0 | 0 | 2 | 2 | 0 | 0 | 0 | 2 | 2 | 0 | 0 | 0 | 2 | 2 |
| Copia-60 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Copia-61 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 1 | 1 |
| Copia-62 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 2 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 |
| Copia-63 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 3 | 3 | 0 | 0 | 0 | 1 | 1 |
| Copia-64 | 0 | 0 | 0 | 2 | 2 | 0 | 0 | 0 | 2 | 2 | 0 | 0 | 2 | 4 | 6 | 0 | 0 | 0 | 4 | 4 |
| Copia-65 | 0 | 1 | 1 | 3 | 5 | 0 | 0 | 1 | 5 | 6 | 0 | 0 | 1 | 3 | 4 | 0 | 0 | 1 | 3 | 4 |
| Copia-66 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 1 | 1 | 0 | 3 | 0 | 3 | 6 | 0 | 0 | 0 | 1 | 1 |
| Copia-67 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 1 | 1 | 2 | 0 | 0 | 2 | 5 | 7 | 0 | 0 | 1 | 0 | 1 |
| Copia-68 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 2 | 2 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 1 | 1 |
| Copia-69 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 1 | 0 | 1 |
| Copia-6 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

| | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Copia-70 | 0 | 8 | 0 | 15 | 23 | 0 | 11 | 0 | 13 | 24 | 0 | 11 | 0 | 12 | 23 | 0 | 10 | 0 | 13 | 23 |
| Copia-71 | 0 | 2 | 0 | 12 | 14 | 0 | 1 | 0 | 16 | 17 | 0 | 1 | 0 | 10 | 11 | 0 | 1 | 0 | 11 | 12 |
| Copia-72 | 0 | 0 | 0 | 5 | 5 | 0 | 0 | 0 | 10 | 10 | 0 | 0 | 0 | 6 | 6 | 0 | 0 | 0 | 7 | 7 |
| Copia-73 | 0 | 0 | 0 | 3 | 3 | 0 | 0 | 0 | 3 | 3 | 0 | 0 | 0 | 2 | 2 | 0 | 0 | 0 | 3 | 3 |
| Copia-74 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Copia-75 | 0 | 0 | 5 | 5 | 10 | 0 | 0 | 6 | 8 | 14 | 0 | 0 | 5 | 8 | 13 | 0 | 0 | 5 | 10 | 15 |
| Copia-76 | 0 | 20 | 6 | 20 | 46 | 2 | 34 | 8 | 24 | 68 | 0 | 16 | 7 | 21 | 44 | 0 | 19 | 5 | 17 | 41 |
| Copia-77 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| Copia-78 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Copia-79 | 0 | 0 | 0 | 2 | 2 | 0 | 0 | 1 | 3 | 4 | 0 | 0 | 0 | 3 | 3 | 0 | 0 | 0 | 5 | 5 |
| Copia-7 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Copia-80 | 0 | 1 | 1 | 17 | 19 | 0 | 5 | 1 | 30 | 36 | 0 | 2 | 1 | 49 | 52 | 0 | 6 | 0 | 52 | 58 |
| Copia-81 | 0 | 0 | 0 | 2 | 2 | 0 | 0 | 0 | 5 | 5 | 0 | 1 | 0 | 6 | 7 | 0 | 0 | 0 | 5 | 5 |
| Copia-82 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 2 | 2 | 0 | 2 | 0 | 6 | 8 | 0 | 0 | 1 | 2 | 3 |
| Copia-83 | 0 | 1 | 0 | 1 | 2 | 0 | 1 | 0 | 2 | 3 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 1 | 2 | 3 |
| Copia-84 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 2 | 2 | 4 | 0 | 0 | 1 | 1 | 2 | 0 | 0 | 0 | 2 | 2 |
| Copia-85 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 2 | 2 | 0 | 0 | 0 | 2 | 2 | 0 | 0 | 0 | 3 | 3 |
| Copia-86 | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 1 | 2 | 0 | 0 | 0 | 2 | 2 | 0 | 0 | 0 | 2 | 2 |
| Copia-87 | 0 | 0 | 0 | 6 | 6 | 0 | 0 | 0 | 8 | 8 | 0 | 0 | 0 | 12 | 12 | 0 | 0 | 0 | 12 | 12 |
| Copia-88 | 0 | 0 | 0 | 2 | 2 | 0 | 0 | 0 | 8 | 8 | 0 | 0 | 0 | 11 | 11 | 0 | 0 | 0 | 5 | 5 |
| Copia-89 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 3 | 3 | 0 | 0 | 0 | 3 | 3 | 0 | 0 | 0 | 3 | 3 |
| Copia-8 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Copia-90 | 0 | 2 | 0 | 4 | 6 | 0 | 1 | 0 | 6 | 7 | 0 | 1 | 0 | 5 | 6 | 0 | 0 | 0 | 6 | 6 |
| Copia-91 | 1 | 3 | 0 | 1 | 5 | 0 | 4 | 0 | 1 | 5 | 0 | 0 | 0 | 3 | 3 | 0 | 0 | 0 | 2 | 2 |
| Copia-92 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Copia-93 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 1 | 2 |
| Copia-94 | 0 | 1 | 9 | 10 | 20 | 0 | 4 | 15 | 20 | 39 | 0 | 2 | 30 | 37 | 69 | 0 | 2 | 22 | 22 | 46 |
| Copia-95 | 0 | 0 | 0 | 3 | 3 | 0 | 0 | 0 | 3 | 3 | 0 | 1 | 0 | 2 | 3 | 0 | 2 | 0 | 2 | 4 |
| Copia-96 | 0 | 1 | 0 | 7 | 8 | 0 | 0 | 0 | 8 | 8 | 0 | 0 | 1 | 5 | 6 | 0 | 0 | 1 | 7 | 8 |
| Copia-97 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| Copia-98 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Copia-99 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 |
| Copia-9 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 | 3 | 0 | 0 | 0 | 6 | 6 | 0 | 0 | 0 | 5 | 5 |
| Copia-Tvv1 | 0 | 0 | 1 | 5 | 6 | 0 | 1 | 1 | 10 | 12 | 2 | 5 | 0 | 11 | 18 | 0 | 0 | 0 | 10 | 10 |

## Table C.2 Gypsy expression candidates grouped by families, distinctness and integrity

| Gypsy Family | T=0 | | | | | Mock | | | | | Yeast | | | | | Botrytis | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Expression candidates | | | | | Expression candidates | | | | | Expression candidates | | | | | Expression candidates | | | | |
| | Untrackable | | Trackable | | Sum | Untrackable | | Trackable | | Sum | Untrackable | | Trackable | | Sum | Untrackable | | Trackable | | Sum |
| | FL | FG | FL | FG | | FL | FG | FL | FG | | FL | FG | FL | FG | | FL | FG | FL | FG | |
| Gypsy-10 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Gypsy-11 | 0 | 13 | 1 | 2 | 16 | 1 | 15 | 2 | 3 | 21 | 0 | 8 | 0 | 2 | 10 | 1 | 6 | 2 | 1 | 10 |
| Gypsy-12 | 0 | 162 | 0 | 13 | 175 | 1 | 135 | 0 | 14 | 150 | 0 | 130 | 0 | 6 | 136 | 0 | 76 | 0 | 5 | 81 |
| Gypsy-13 | 0 | 70 | 0 | 14 | 84 | 0 | 64 | 0 | 12 | 76 | 0 | 32 | 0 | 7 | 39 | 0 | 19 | 0 | 8 | 27 |
| Gypsy-14 | 0 | 22 | 0 | 2 | 24 | 0 | 23 | 0 | 0 | 23 | 1 | 30 | 1 | 5 | 37 | 0 | 24 | 1 | 0 | 25 |
| Gypsy-15 | 0 | 1 | 1 | 0 | 2 | 0 | 1 | 1 | 0 | 2 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 1 | 1 | 2 |
| Gypsy-16 | 0 | 7 | 0 | 4 | 11 | 0 | 6 | 0 | 5 | 11 | 0 | 1 | 0 | 7 | 8 | 0 | 2 | 0 | 6 | 8 |
| Gypsy-17 | 0 | 2 | 0 | 2 | 4 | 0 | 2 | 0 | 3 | 5 | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 2 | 3 |
| Gypsy-18 | 1 | 17 | 1 | 20 | 39 | 0 | 4 | 0 | 20 | 24 | 0 | 6 | 2 | 20 | 28 | 0 | 5 | 2 | 15 | 22 |
| Gypsy-19 | 0 | 39 | 0 | 14 | 53 | 0 | 33 | 0 | 14 | 47 | 0 | 29 | 1 | 11 | 41 | 0 | 20 | 1 | 14 | 35 |
| Gypsy-1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 3 | 3 | 0 | 0 | 0 | 0 | 0 |
| Gypsy-20 | 1 | 10 | 0 | 2 | 13 | 1 | 7 | 0 | 2 | 10 | 0 | 2 | 0 | 2 | 4 | 0 | 3 | 0 | 2 | 5 |
| Gypsy-21 | 0 | 28 | 0 | 9 | 37 | 0 | 8 | 0 | 12 | 20 | 0 | 5 | 0 | 15 | 20 | 0 | 7 | 0 | 11 | 18 |
| Gypsy-22 | 0 | 1 | 0 | 5 | 6 | 0 | 1 | 0 | 5 | 6 | 0 | 0 | 0 | 2 | 2 | 0 | 0 | 0 | 3 | 3 |
| Gypsy-23 | 0 | 2 | 0 | 21 | 23 | 0 | 0 | 0 | 28 | 28 | 0 | 2 | 0 | 26 | 28 | 0 | 1 | 0 | 25 | 26 |
| Gypsy-24 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Gypsy-26 | 0 | 32 | 0 | 49 | 81 | 0 | 33 | 0 | 66 | 99 | 0 | 43 | 0 | 88 | 131 | 0 | 33 | 0 | 87 | 120 |
| Gypsy-27 | 0 | 0 | 0 | 6 | 6 | 0 | 3 | 0 | 4 | 7 | 0 | 0 | 0 | 17 | 17 | 0 | 1 | 0 | 10 | 11 |
| Gypsy-28 | 0 | 4 | 0 | 7 | 11 | 0 | 4 | 0 | 9 | 13 | 0 | 3 | 0 | 9 | 12 | 0 | 4 | 0 | 9 | 13 |
| Gypsy-29 | 0 | 1 | 0 | 6 | 7 | 0 | 2 | 0 | 6 | 8 | 0 | 0 | 0 | 3 | 3 | 0 | 1 | 0 | 3 | 4 |
| Gypsy-2 | 0 | 1 | 0 | 10 | 11 | 0 | 5 | 0 | 13 | 18 | 0 | 1 | 0 | 19 | 20 | 0 | 3 | 0 | 15 | 18 |
| Gypsy-30 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Gypsy-31 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Gypsy-32 | 0 | 0 | 1 | 3 | 4 | 0 | 0 | 1 | 4 | 5 | 0 | 0 | 1 | 3 | 4 | 0 | 0 | 1 | 6 | 7 |
| Gypsy-33 | 0 | 0 | 0 | 5 | 5 | 0 | 1 | 0 | 12 | 13 | 0 | 2 | 1 | 16 | 19 | 0 | 0 | 0 | 14 | 14 |
| Gypsy-34 | 0 | 0 | 0 | 6 | 6 | 0 | 0 | 0 | 9 | 9 | 0 | 0 | 0 | 7 | 7 | 0 | 0 | 0 | 6 | 6 |
| Gypsy-3 | 0 | 0 | 0 | 3 | 3 | 0 | 0 | 0 | 7 | 7 | 0 | 0 | 1 | 11 | 12 | 0 | 0 | 0 | 10 | 10 |
| Gypsy-4 | 0 | 22 | 0 | 2 | 24 | 0 | 32 | 0 | 2 | 34 | 0 | 25 | 0 | 4 | 29 | 0 | 22 | 0 | 2 | 24 |
| Gypsy-5 | 0 | 0 | 0 | 1 | 1 | 0 | 2 | 0 | 1 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 |
| Gypsy-6 | 0 | 4 | 0 | 0 | 4 | 0 | 3 | 0 | 0 | 3 | 0 | 1 | 0 | 2 | 3 | 0 | 1 | 0 | 1 | 2 |
| Gypsy-7 | 0 | 51 | 0 | 6 | 57 | 0 | 76 | 0 | 7 | 83 | 0 | 39 | 0 | 6 | 45 | 0 | 56 | 0 | 5 | 61 |
| Gypsy-8 | 0 | 1 | 0 | 2 | 3 | 0 | 1 | 0 | 2 | 3 | 0 | 0 | 0 | 3 | 3 | 0 | 0 | 0 | 2 | 2 |
| Gypsy-9 | 0 | 2 | 0 | 32 | 34 | 0 | 4 | 1 | 40 | 45 | 0 | 0 | 3 | 42 | 45 | 1 | 2 | 2 | 39 | 44 |
| Gret1 | 0 | 0 | 1 | 1 | 2 | 0 | 1 | 1 | 3 | 5 | 0 | 0 | 2 | 2 | 4 | 0 | 0 | 1 | 1 | 2 |
| GYVIT1 | 3 | 28 | 0 | 4 | 35 | 3 | 34 | 0 | 3 | 40 | 3 | 16 | 0 | 5 | 24 | 2 | 21 | 0 | 4 | 27 |

| | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Gypsy-V1 | 0 | 3 | 0 | 6 | 9 | 3 | 5 | 0 | 9 | 17 | 0 | 0 | 2 | 6 | 8 | 0 | 0 | 1 | 4 | 5 |

## Table C.3 Calimovirus expression candidates grouped by families, distinctness and integrity

| Calimovirus Family | T=0 | | | | | Mock | | | | | Yeast | | | | | Botrytis | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Expression candidates | | | | | Expression candidates | | | | | Expression candidates | | | | | Expression candidates | | | | |
| | Untrackable | | Trackable | | Sum | Untrackable | | Trackable | | Sum | Untrackable | | Trackable | | Sum | Untrackable | | Trackable | | Sum |
| | FL | FG | FL | FG | | FL | FG | FL | FG | | FL | FG | FL | FG | | FL | FG | FL | FG | |
| Caulimovirus-1 | 0 | 1 | 0 | 7 | 8 | 0 | 0 | 0 | 7 | 7 | 0 | 0 | 0 | 9 | 9 | 0 | 0 | 0 | 9 | 9 |
| Caulimovirus-2 | 0 | 2 | 2 | 4 | 8 | 0 | 1 | 2 | 5 | 8 | 0 | 1 | 2 | 7 | 10 | 0 | 1 | 2 | 7 | 10 |
| CAULIV11 | 3 | 14 | 1 | 11 | 29 | 0 | 2 | 1 | 8 | 11 | 0 | 2 | 3 | 10 | 15 | 0 | 2 | 3 | 10 | 15 |

## Table C.4 LINE expression candidates grouped by families, distinctness and integrity

| LINE Family | T=0 | | | | | Mock | | | | | Yeast | | | | | Botrytis | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Expression candidates | | | | | Expression candidates | | | | | Expression candidates | | | | | Expression candidates | | | | |
| | Untrackable | | Trackable | | Sum | Untrackable | | Trackable | | Sum | Untrackable | | Trackable | | Sum | Untrackable | | Trackable | | Sum |
| | FL | FG | FL | FG | | FL | FG | FL | FG | | FL | FG | FL | FG | | FL | FG | FL | FG | |
| VLINE10 | 0 | 0 | 0 | 79 | 79 | 0 | 4 | 1 | 123 | 128 | 0 | 2 | 3 | 113 | 118 | 0 | 2 | 1 | 128 | 131 |
| VLINE1 | 0 | 17 | 1 | 227 | 245 | 0 | 26 | 2 | 409 | 437 | 0 | 22 | 8 | 499 | 529 | 0 | 18 | 7 | 471 | 496 |
| VLINE2 | 0 | 2 | 2 | 71 | 75 | 0 | 9 | 6 | 126 | 141 | 0 | 8 | 20 | 223 | 251 | 0 | 4 | 17 | 185 | 206 |
| VLINE3 | 0 | 2 | 9 | 55 | 66 | 0 | 15 | 20 | 152 | 187 | 0 | 9 | 38 | 217 | 264 | 0 | 10 | 38 | 202 | 250 |
| VLINE4 | 0 | 5 | 0 | 206 | 211 | 0 | 13 | 0 | 349 | 362 | 0 | 13 | 1 | 335 | 349 | 0 | 10 | 1 | 332 | 343 |
| VLINE5 | 0 | 7 | 0 | 110 | 117 | 0 | 17 | 0 | 172 | 189 | 0 | 11 | 0 | 225 | 236 | 0 | 16 | 0 | 195 | 211 |
| VLINE6 | 0 | 22 | 0 | 117 | 139 | 0 | 37 | 0 | 205 | 242 | 0 | 25 | 0 | 281 | 306 | 0 | 26 | 0 | 254 | 280 |
| VLINE7 | 0 | 1 | 2 | 71 | 74 | 0 | 22 | 3 | 148 | 173 | 0 | 12 | 4 | 144 | 160 | 0 | 8 | 3 | 141 | 152 |
| VLINE8 | 0 | 12 | 6 | 53 | 71 | 0 | 34 | 12 | 136 | 182 | 0 | 26 | 14 | 189 | 229 | 0 | 24 | 15 | 181 | 220 |
| VLINE9 | 0 | 4 | 1 | 47 | 52 | 0 | 11 | 4 | 75 | 90 | 0 | 5 | 8 | 66 | 79 | 0 | 4 | 5 | 68 | 77 |

## Table C.5 CACTA expression candidates grouped by families, distinctness and integrity

| CACTA Family | T=0 | | | | | Mock | | | | | Yeast | | | | | Botrytis | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Expression candidates | | | | | Expression candidates | | | | | Expression candidates | | | | | Expression candidates | | | | |
| | Untrackable | | Trackable | | Sum | Untrackable | | Trackable | | Sum | Untrackable | | Trackable | | Sum | Untrackable | | Trackable | | Sum |
| | FL | FG | FL | FG | | FL | FG | FL | FG | | FL | FG | FL | FG | | FL | FG | FL | FG | |
| CACTA-10 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| CACTA-11 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 1 | 0 | 1 | 2 | 0 | 1 | 0 | 0 | 1 |
| CACTA-12 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| CACTA-13 | 0 | 7 | 0 | 24 | 31 | 0 | 12 | 0 | 23 | 35 | 0 | 8 | 0 | 14 | 22 | 0 | 7 | 0 | 17 | 24 |
| CACTA-1 | 0 | 0 | 0 | 4 | 4 | 0 | 0 | 0 | 4 | 4 | 0 | 0 | 0 | 4 | 4 | 0 | 0 | 0 | 4 | 4 |
| CACTA-2 | 0 | 3 | 0 | 4 | 7 | 0 | 3 | 0 | 9 | 12 | 0 | 3 | 0 | 5 | 8 | 0 | 3 | 0 | 6 | 9 |
| CACTA-3 | 0 | 0 | 0 | 13 | 13 | 0 | 0 | 0 | 16 | 16 | 0 | 0 | 0 | 13 | 13 | 0 | 0 | 0 | 14 | 14 |
| CACTA-4 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 2 | 2 | 0 | 0 | 0 | 2 | 2 | 0 | 0 | 0 | 2 | 2 |
| CACTA-4N1 | 0 | 0 | 0 | 1 | 1 | 2 | 2 | 0 | 2 | 6 | 1 | 1 | 0 | 1 | 3 | 0 | 0 | 0 | 2 | 2 |
| CACTA-5 | 0 | 6 | 0 | 18 | 24 | 0 | 9 | 0 | 23 | 32 | 0 | 4 | 0 | 16 | 20 | 0 | 7 | 0 | 17 | 24 |
| CACTA-6 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 2 | 2 | 0 | 0 | 0 | 2 | 2 | 0 | 0 | 0 | 1 | 1 |
| CACTA-7 | 0 | 3 | 0 | 11 | 14 | 0 | 3 | 0 | 15 | 18 | 0 | 2 | 0 | 14 | 16 | 0 | 2 | 0 | 12 | 14 |
| CACTA-8N | 0 | 0 | 0 | 1 | 1 | 2 | 0 | 0 | 3 | 5 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 1 | 1 |
| CACTA-9 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| CACTA-N3 | 0 | 0 | 0 | 2 | 2 | 0 | 0 | 0 | 7 | 7 | 0 | 0 | 0 | 3 | 3 | 0 | 1 | 0 | 4 | 5 |

## Table C.6 Harbinger expression candidates grouped by families, distinctness and integrity

| Harbinger Family | T=0 | | | | | Mock | | | | | Yeast | | | | | Botrytis | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Expression candidates | | | | | Expression candidates | | | | | Expression candidates | | | | | Expression candidates | | | | |
| | Untrackable | | Trackable | | Sum | Untrackable | | Trackable | | Sum | Untrackable | | Trackable | | Sum | Untrackable | | Trackable | | Sum |
| | FL | FG | FL | FG | | FL | FG | FL | FG | | FL | FG | FL | FG | | FL | FG | FL | FG | |
| Harbinger-1 | 0 | 3 | 0 | 17 | 20 | 0 | 4 | 0 | 19 | 23 | 0 | 0 | 0 | 15 | 15 | 0 | 0 | 0 | 15 | 15 |
| Harbinger-1N1 | 0 | 0 | 0 | 18 | 18 | 0 | 1 | 1 | 21 | 23 | 0 | 1 | 1 | 13 | 15 | 0 | 0 | 0 | 17 | 17 |
| Harbinger-3 | 0 | 5 | 0 | 16 | 21 | 0 | 8 | 0 | 22 | 30 | 0 | 4 | 0 | 19 | 23 | 0 | 5 | 0 | 21 | 26 |
| Harbinger-3N1 | 0 | 0 | 10 | 4 | 14 | 0 | 0 | 13 | 7 | 20 | 1 | 1 | 8 | 3 | 13 | 0 | 0 | 7 | 5 | 12 |
| Harbinger-3N2 | 0 | 1 | 1 | 4 | 6 | 0 | 0 | 2 | 4 | 6 | 0 | 0 | 0 | 2 | 2 | 0 | 1 | 2 | 3 | 6 |
| Harbinger-3N3B | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 3 | 3 | 0 | 0 | 0 | 2 | 2 | 0 | 0 | 0 | 2 | 2 |
| Harbinger-3N3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| VHARB4 | 0 | 4 | 0 | 39 | 43 | 0 | 3 | 0 | 51 | 54 | 0 | 4 | 0 | 40 | 44 | 0 | 3 | 0 | 40 | 43 |
| VHARB-N1 | 0 | 1 | 33 | 63 | 97 | 0 | 0 | 49 | 80 | 129 | 0 | 2 | 26 | 49 | 77 | 0 | 2 | 32 | 54 | 88 |
| VHARB-N2 | 0 | 4 | 0 | 24 | 28 | 0 | 7 | 0 | 27 | 34 | 0 | 4 | 0 | 17 | 21 | 0 | 4 | 0 | 19 | 23 |
| VHARB-N3 | 0 | 0 | 7 | 28 | 35 | 0 | 1 | 9 | 49 | 59 | 0 | 2 | 4 | 28 | 34 | 0 | 1 | 5 | 34 | 40 |
| Harbinger-1 | 0 | 3 | 0 | 17 | 20 | 0 | 4 | 0 | 19 | 23 | 0 | 0 | 0 | 15 | 15 | 0 | 0 | 0 | 15 | 15 |

| | | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Harbinger-1N1 | 0 | 0 | 0 | 18 | 18 | 0 | 1 | 1 | 21 | 23 | 0 | 1 | 1 | 13 | 15 | 0 | 0 | 0 | 17 | 17 |
| CACTA-9 | 0 | 5 | 0 | 16 | 21 | 0 | 8 | 0 | 22 | 30 | 0 | 4 | 0 | 19 | 23 | 0 | 5 | 0 | 21 | 26 |
| CACTA-N3 | 0 | 0 | 10 | 4 | 14 | 0 | 0 | 13 | 7 | 20 | 1 | 1 | 8 | 3 | 13 | 0 | 0 | 7 | 5 | 12 |

## Table C.7 hAT expression candidates grouped by families, distinctness and integrity

| hAT Family | T=0 | | | | | Mock | | | | | Yeast | | | | | Botrytis | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Expression candidates | | | | | Expression candidates | | | | | Expression candidates | | | | | Expression candidates | | | | |
| | Untrackable | | Trackable | | Sum | Untrackable | | Trackable | | Sum | Untrackable | | Trackable | | Sum | Untrackable | | Trackable | | Sum |
| | FL | FG | FL | FG | | FL | FG | FL | FG | | FL | FG | FL | FG | | FL | FG | FL | FG | |
| hAT-10 | 0 | 0 | 0 | 7 | 7 | 0 | 1 | 0 | 7 | 8 | 0 | 0 | 0 | 6 | 6 | 0 | 1 | 0 | 6 | 7 |
| hAT-10N1 | 0 | 0 | 0 | 8 | 8 | 0 | 7 | 0 | 21 | 28 | 0 | 2 | 0 | 12 | 14 | 0 | 7 | 0 | 9 | 16 |
| hAT-11N | 0 | 0 | 0 | 2 | 2 | 0 | 0 | 0 | 4 | 4 | 0 | 0 | 0 | 8 | 8 | 0 | 0 | 0 | 6 | 6 |
| hAT-12 | 1 | 11 | 0 | 15 | 27 | 0 | 12 | 0 | 15 | 27 | 0 | 8 | 0 | 13 | 21 | 0 | 10 | 0 | 14 | 24 |
| hAT-13 | 0 | 0 | 0 | 6 | 6 | 0 | 0 | 0 | 6 | 6 | 0 | 1 | 1 | 7 | 9 | 0 | 0 | 1 | 7 | 8 |
| hAT-6 | 0 | 0 | 0 | 19 | 19 | 0 | 0 | 0 | 20 | 20 | 0 | 0 | 0 | 19 | 19 | 0 | 2 | 0 | 21 | 23 |
| hAT-7 | 4 | 16 | 0 | 16 | 36 | 4 | 21 | 1 | 18 | 44 | 2 | 11 | 1 | 15 | 29 | 2 | 8 | 0 | 14 | 24 |
| TE-7-1 | 0 | 2 | 0 | 4 | 6 | 0 | 1 | 3 | 11 | 15 | 0 | 4 | 1 | 4 | 9 | 0 | 2 | 2 | 6 | 10 |
| VIHAT1 | 0 | 0 | 1 | 13 | 14 | 5 | 27 | 1 | 21 | 54 | 6 | 33 | 5 | 22 | 66 | 6 | 28 | 2 | 24 | 60 |
| VIHAT2 | 0 | 3 | 0 | 16 | 19 | 0 | 4 | 0 | 25 | 29 | 0 | 2 | 0 | 15 | 17 | 0 | 4 | 0 | 20 | 24 |
| VIHAT2-N1 | 0 | 0 | 1 | 3 | 4 | 0 | 0 | 1 | 8 | 9 | 0 | 1 | 0 | 3 | 4 | 0 | 0 | 1 | 5 | 6 |
| VIHAT3 | 4 | 16 | 2 | 21 | 43 | 7 | 20 | 2 | 31 | 60 | 7 | 13 | 4 | 22 | 46 | 2 | 11 | 3 | 28 | 44 |
| VIHAT3-N1 | 0 | 1 | 0 | 9 | 10 | 0 | 1 | 0 | 11 | 12 | 0 | 2 | 0 | 8 | 10 | 0 | 0 | 0 | 10 | 10 |
| Vinesleeper-1 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 1 | 1 |
| Vinesleeper-2 | 0 | 0 | 1 | 3 | 4 | 0 | 0 | 1 | 3 | 4 | 0 | 0 | 1 | 3 | 4 | 0 | 0 | 1 | 3 | 4 |

## Table C.8 Helitron expression candidates grouped by families, distinctness and integrity

| Helitron Family | T=0 | | | | | Mock | | | | | Yeast | | | | | Botrytis | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Expression candidates | | | | | Expression candidates | | | | | Expression candidates | | | | | Expression candidates | | | | |
| | Untrackable | | Trackable | | Sum | Untrackable | | Trackable | | Sum | Untrackable | | Trackable | | Sum | Untrackable | | Trackable | | Sum |
| | FL | FG | FL | FG | | FL | FG | FL | FG | | FL | FG | FL | FG | | FL | FG | FL | FG | |
| Helitron-1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

## Table C.9 MULE expression candidates grouped by families, distinctness and integrity

| MULE Family | T=0 | | | | | Mock | | | | | Yeast | | | | | Botrytis | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Expression candidates | | | | | Expression candidates | | | | | Expression candidates | | | | | Expression candidates | | | | |
| | Untrackable | | Trackable | | Sum | Untrackable | | Trackable | | Sum | Untrackable | | Trackable | | Sum | Untrackable | | Trackable | | Sum |
| | FL | FG | FL | FG | | FL | FG | FL | FG | | FL | FG | FL | FG | | FL | FG | FL | FG | |
| Hopvine-1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Hopvine-2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Jithouse-1 | 0 | 0 | 1 | 6 | 7 | 0 | 1 | 1 | 5 | 7 | 0 | 0 | 1 | 6 | 7 | 0 | 0 | 1 | 5 | 6 |
| Jithouse-2 | 0 | 0 | 1 | 10 | 11 | 0 | 0 | 1 | 11 | 12 | 0 | 0 | 1 | 11 | 12 | 0 | 0 | 1 | 11 | 12 |
| Jithouse-3 | 0 | 6 | 0 | 50 | 56 | 0 | 4 | 0 | 73 | 77 | 0 | 6 | 0 | 46 | 52 | 0 | 3 | 0 | 53 | 56 |
| Jithouse-4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Jithouse-5 | 0 | 0 | 2 | 12 | 14 | 0 | 0 | 2 | 12 | 14 | 0 | 0 | 2 | 12 | 14 | 0 | 0 | 2 | 12 | 14 |
| Jitvine-2 | 0 | 14 | 0 | 13 | 27 | 0 | 12 | 0 | 23 | 35 | 0 | 6 | 0 | 15 | 21 | 0 | 9 | 0 | 9 | 18 |
| MuDR-11N | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| MuDR-12 | 0 | 11 | 0 | 27 | 38 | 0 | 21 | 0 | 29 | 50 | 0 | 9 | 0 | 13 | 22 | 0 | 5 | 0 | 17 | 22 |
| MuDR-13 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 |
| MuDR-13NB | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| MuDR-13NC | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| MuDR-18 | 0 | 0 | 0 | 2 | 2 | 0 | 0 | 0 | 2 | 2 | 0 | 0 | 0 | 2 | 2 | 0 | 0 | 0 | 2 | 2 |
| MuDR-21 | 0 | 0 | 0 | 10 | 10 | 0 | 0 | 0 | 12 | 12 | 0 | 0 | 0 | 9 | 9 | 0 | 0 | 0 | 11 | 11 |
| MuDR-22 | 0 | 0 | 0 | 13 | 13 | 0 | 0 | 0 | 18 | 18 | 0 | 0 | 0 | 8 | 8 | 0 | 0 | 0 | 11 | 11 |
| MuDR-3 | 0 | 0 | 0 | 4 | 4 | 0 | 0 | 0 | 4 | 4 | 0 | 0 | 0 | 5 | 5 | 0 | 0 | 0 | 4 | 4 |
| MuDR-4 | 0 | 0 | 0 | 4 | 4 | 0 | 0 | 0 | 4 | 4 | 0 | 2 | 0 | 4 | 6 | 0 | 0 | 0 | 2 | 2 |
| MuDR-5 | 0 | 1 | 0 | 9 | 10 | 0 | 0 | 0 | 11 | 11 | 0 | 0 | 0 | 10 | 10 | 0 | 0 | 0 | 8 | 8 |
| MuDR-6 | 0 | 1 | 0 | 24 | 25 | 0 | 1 | 0 | 30 | 31 | 0 | 0 | 0 | 24 | 24 | 0 | 2 | 0 | 22 | 24 |
| MuDR-7 | 0 | 0 | 0 | 8 | 8 | 0 | 0 | 0 | 9 | 9 | 0 | 0 | 1 | 4 | 5 | 0 | 0 | 0 | 5 | 5 |
| MuDR-8 | 0 | 0 | 0 | 10 | 10 | 0 | 0 | 0 | 12 | 12 | 0 | 4 | 0 | 7 | 11 | 0 | 0 | 0 | 8 | 8 |
| MuDR-9 | 0 | 2 | 0 | 4 | 6 | 0 | 1 | 0 | 5 | 6 | 0 | 1 | 0 | 3 | 4 | 0 | 1 | 0 | 3 | 4 |
| MUDRAVI1 | 0 | 2 | 0 | 34 | 36 | 0 | 5 | 0 | 36 | 41 | 0 | 2 | 0 | 32 | 34 | 0 | 4 | 0 | 28 | 32 |
| MUDRAVI2 | 0 | 0 | 0 | 18 | 18 | 0 | 0 | 0 | 27 | 27 | 0 | 1 | 0 | 16 | 17 | 0 | 2 | 0 | 18 | 20 |
| MuDR-N1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 |
| MUGvine-1 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 1 | 0 | 1 |
| MUGvine-2 | 0 | 0 | 1 | 1 | 2 | 0 | 0 | 1 | 1 | 2 | 0 | 0 | 1 | 1 | 2 | 0 | 0 | 1 | 1 | 2 |
| MUGvine-3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 1 | 0 | 1 |
| MUGvine-4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 1 | 0 | 1 |

| | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| MUGvine-5 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| MUGvine-6 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| MUGvine-7 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| MUGvine-8 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Mutavine-10 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Mutavine-14 | 0 | 2 | 0 | 11 | 13 | 0 | 3 | 0 | 24 | 27 | 0 | 0 | 0 | 17 | 17 | 0 | 0 | 0 | 18 | 18 |
| Mutavine-15 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Mutavine-16 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Mutavine-17 | 0 | 4 | 2 | 40 | 46 | 1 | 5 | 1 | 48 | 55 | 0 | 8 | 1 | 32 | 41 | 0 | 5 | 1 | 34 | 40 |
| Mutavine-18 | 0 | 4 | 1 | 15 | 20 | 0 | 5 | 1 | 18 | 24 | 0 | 1 | 1 | 14 | 16 | 0 | 1 | 1 | 15 | 17 |

## C.4 Test for sequencing reads contributed from four groups of expression candidates: untrackable fragmented, untrackable full-length, trackable full-length, trackable fragmented.

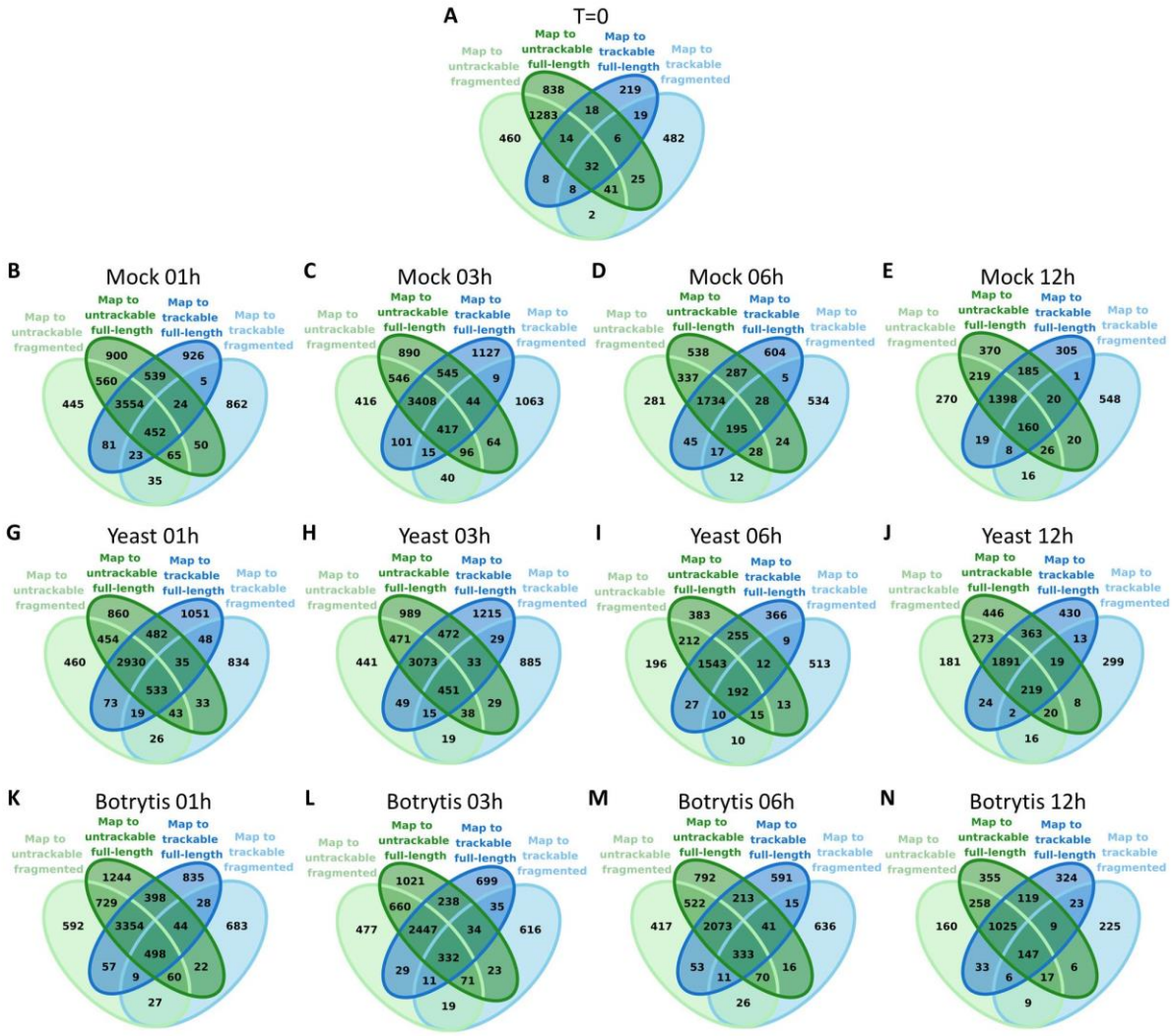



**Figure C. 3 Grouping reads mapping to Copia-3 expression candidates**

Reads mapping to Copia-3 expression candidates were categorized into four groups as indicated. Replicates of each time point were combined.

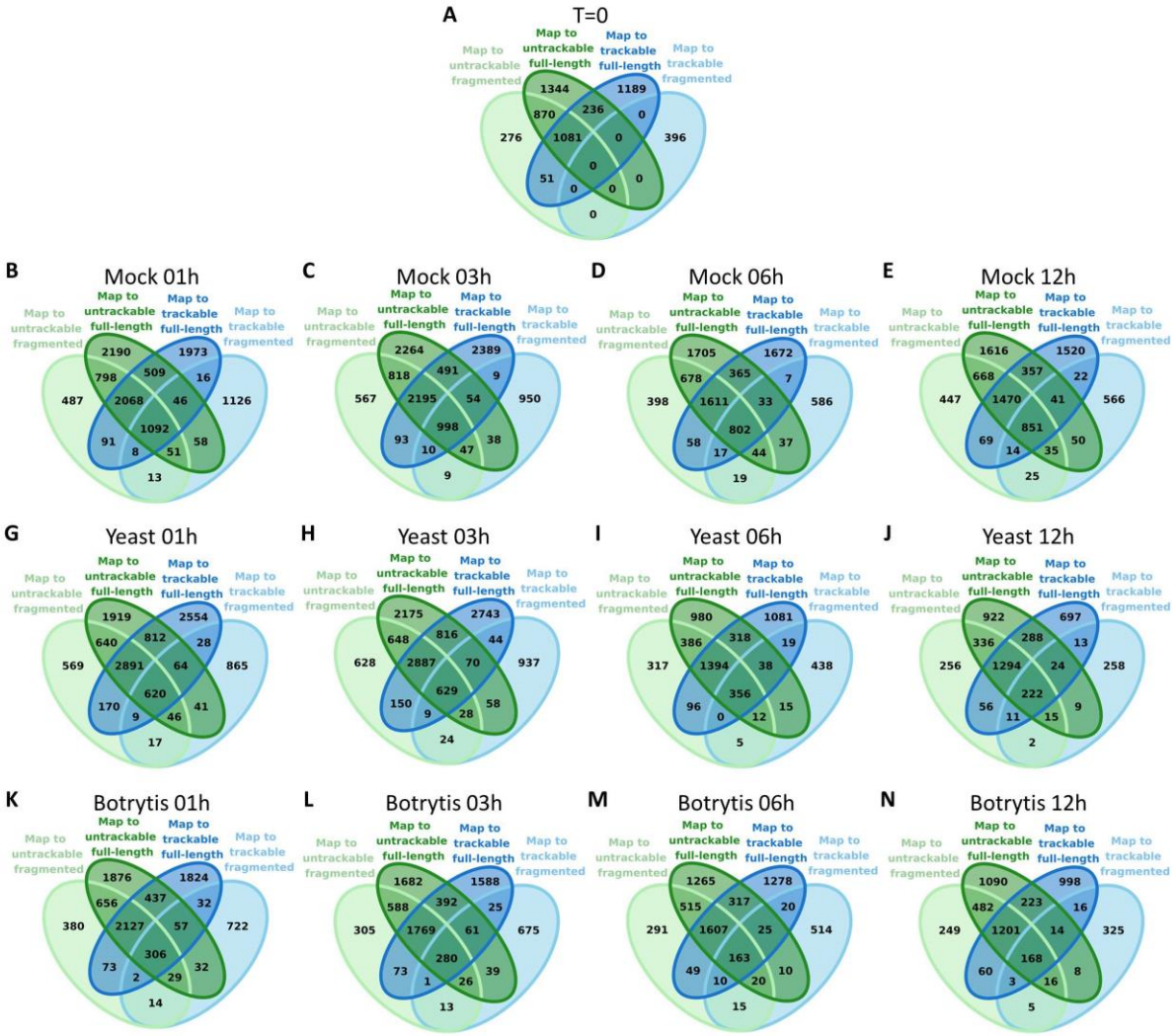



**Figure C.4 Grouping reads mapping to Copia-23 expression candidates**

Reads mapping to Copia-23 expression candidates were categorized into four groups as indicated. Replicates of each time point were combined.

## C.5 Genic and intergenic distribution of annotated TEs and expression candidates from superfamilies contributed to the majority of expression candidates.

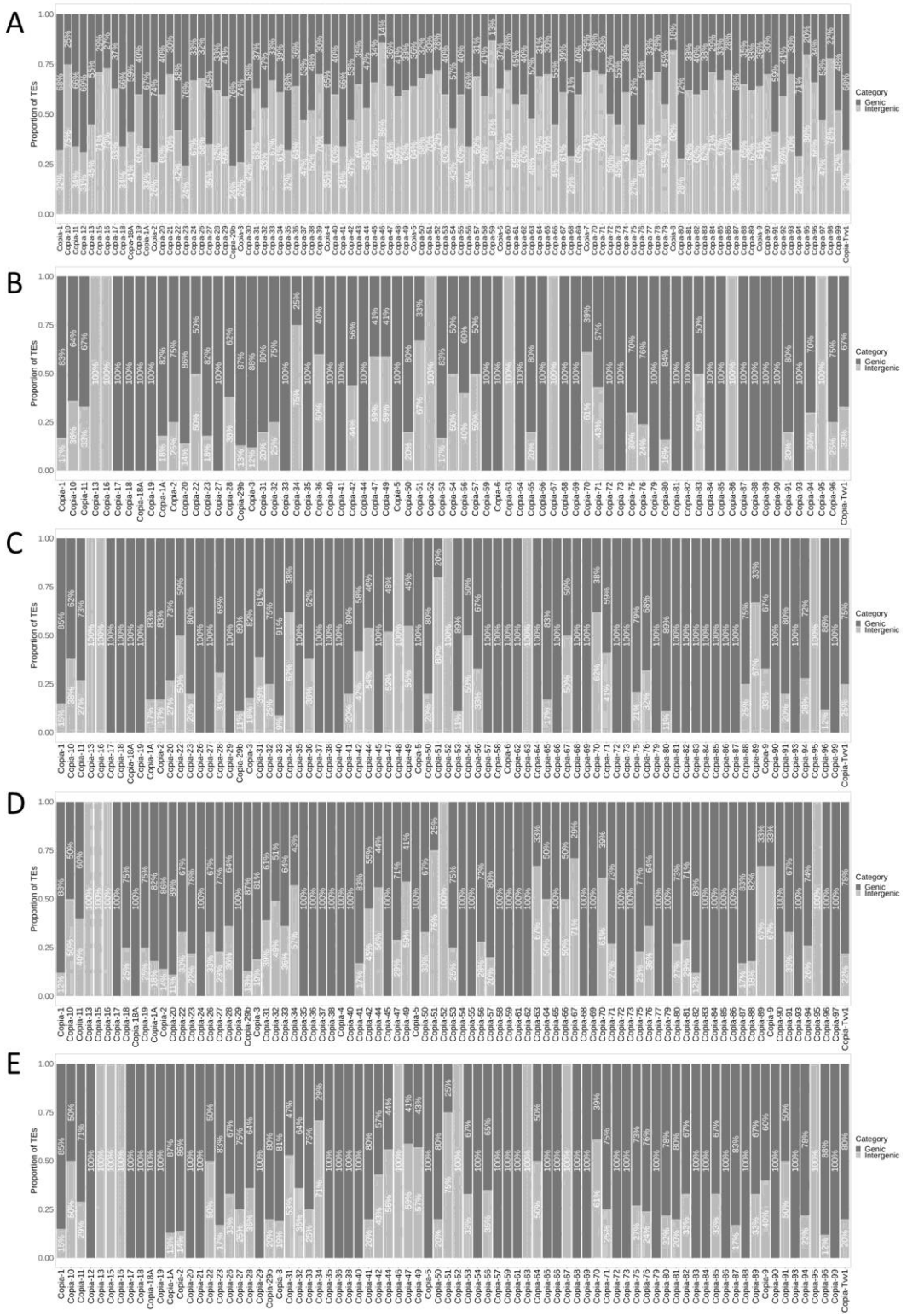



**Figure C.5 Genic and intergenic distribution of annotated TEs and expression candidates of Copia**

**(A)** Distribution of all annotated Copia in the reference genome. **(B-E)** Distribution of Copia expression candidates of T=0 (B), mock (C), yeast (D), and *Botrytis* (E).

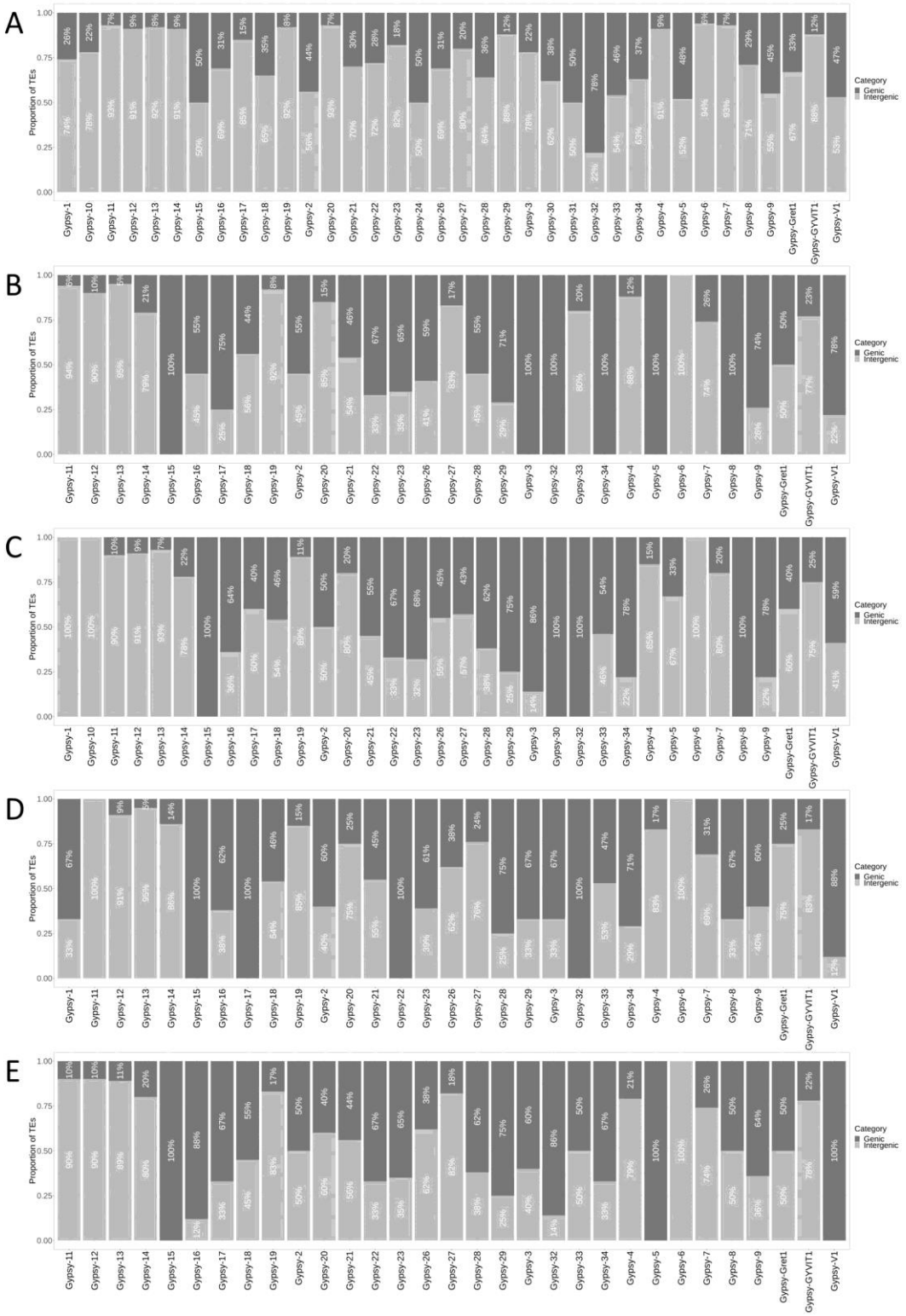



**Figure C.6 Genic and intergenic distribution of annotated TEs and expression candidates of Gypsy**

**(A)** Distribution of all annotated Gypsy in the reference genome. **(B-E)** Distribution of Gypsy expression candidates of T=0 (B), mock (C), yeast (D), and *Botrytis* (E).

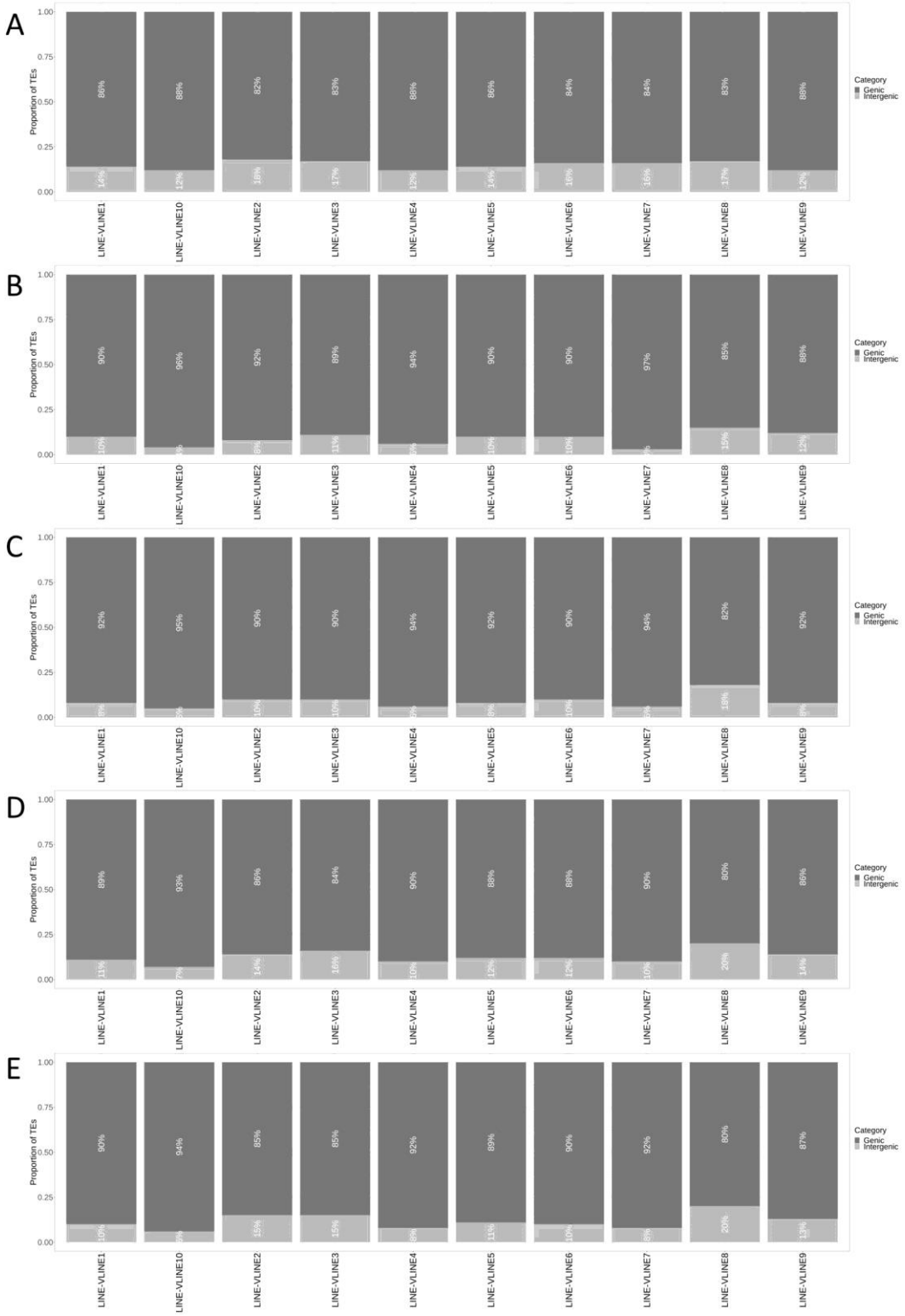



**Figure C.7 Genic and intergenic distribution of annotated TEs and expression candidates of LINE**

**(A)** Distribution of all annotated LINE in the reference genome. **(B-E)** Distribution of LINE expression candidates of T=0 (B), mock (C), yeast (D), and *Botrytis* (E).

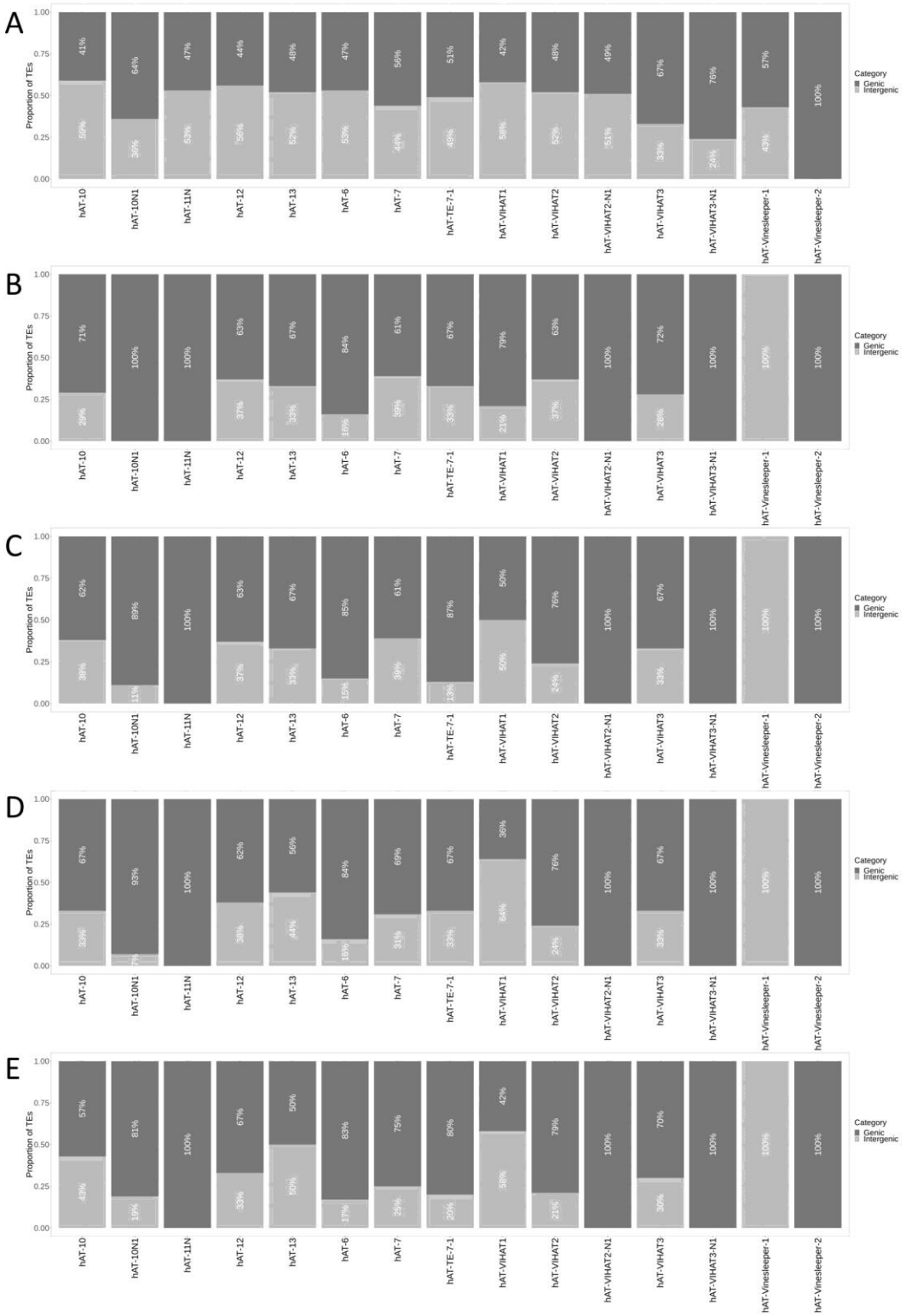



**Figure C.8 Genic and intergenic distribution of annotated TEs and expression candidates of hAT**

**(A)** Distribution of all annotated hAT in the reference genome. **(B-E)** Distribution of hAT expression candidates of T=0 (B), mock (C), yeast (D), and *Botrytis* (E).

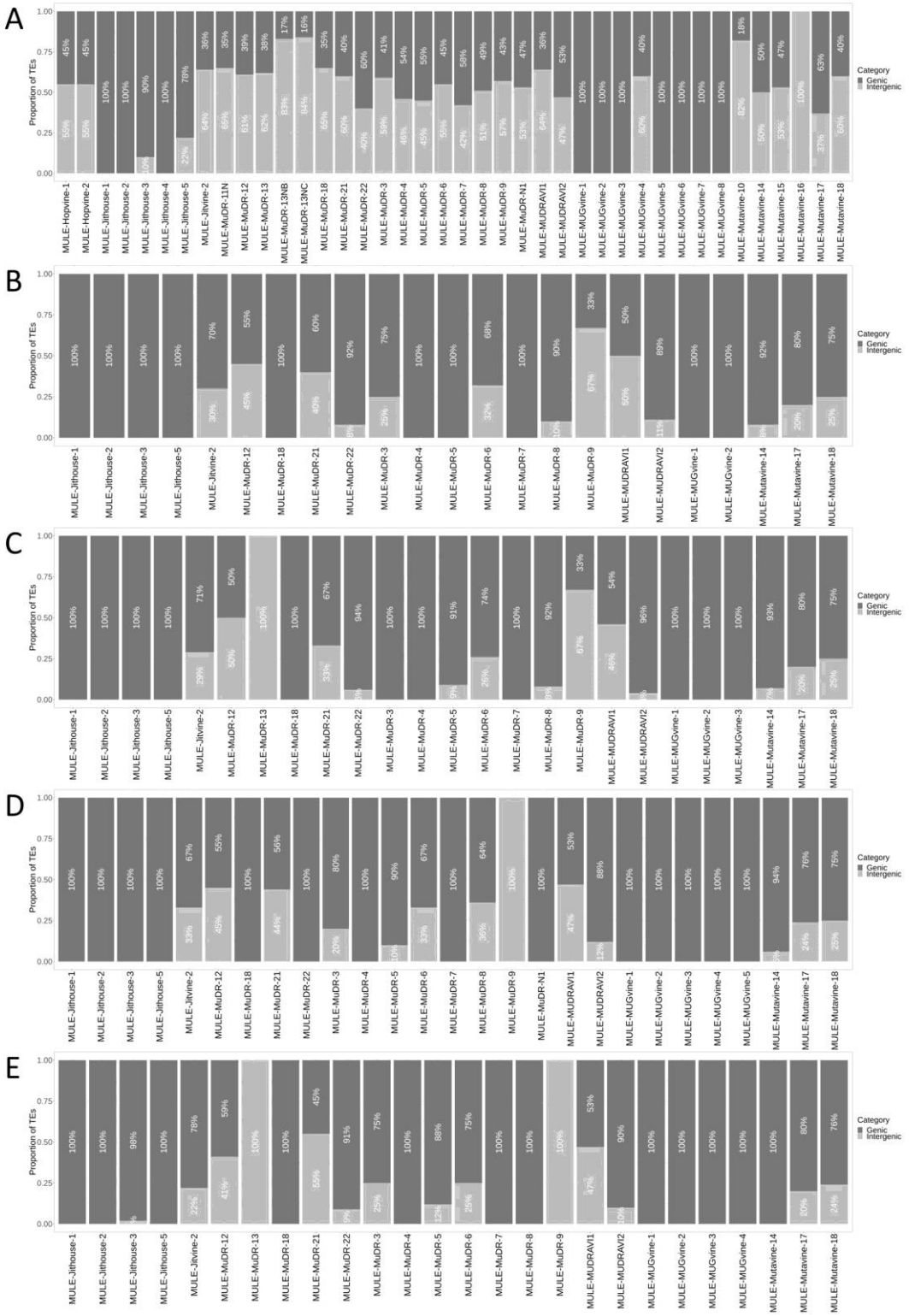



**Figure C.9 Genic and intergenic distribution of annotated TEs and expression candidates of MULE**

**(A)** Distribution of all annotated MULE in the reference genome. **(B-E)** Distribution of MULE expression candidates of T=0 (B), mock (C), yeast (D), and *Botrytis* (E).

## C.6 Location distribution of genic annotated TEs and expression candidates from superfamilies contributed to the majority of expression candidates

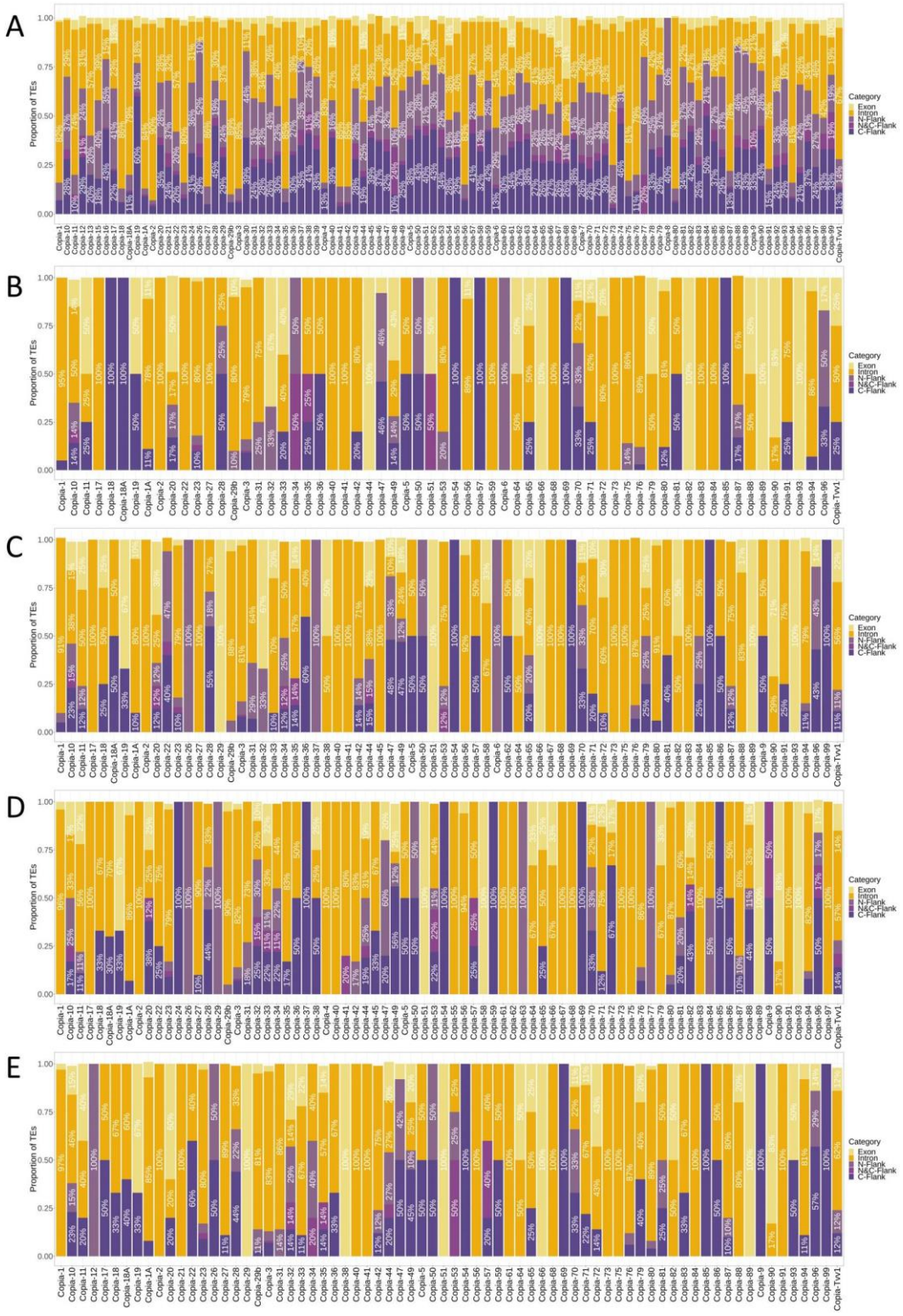



**Figure C.10 Location distribution of annotated genic TEs and expression candidates of Copia**

(A) Location distribution of all annotated genic Copia in the reference genome. (B-E) Location distribution of genic Copia expression candidates of T=0 (B), mock (C), yeast (D), and Botrytis (E).

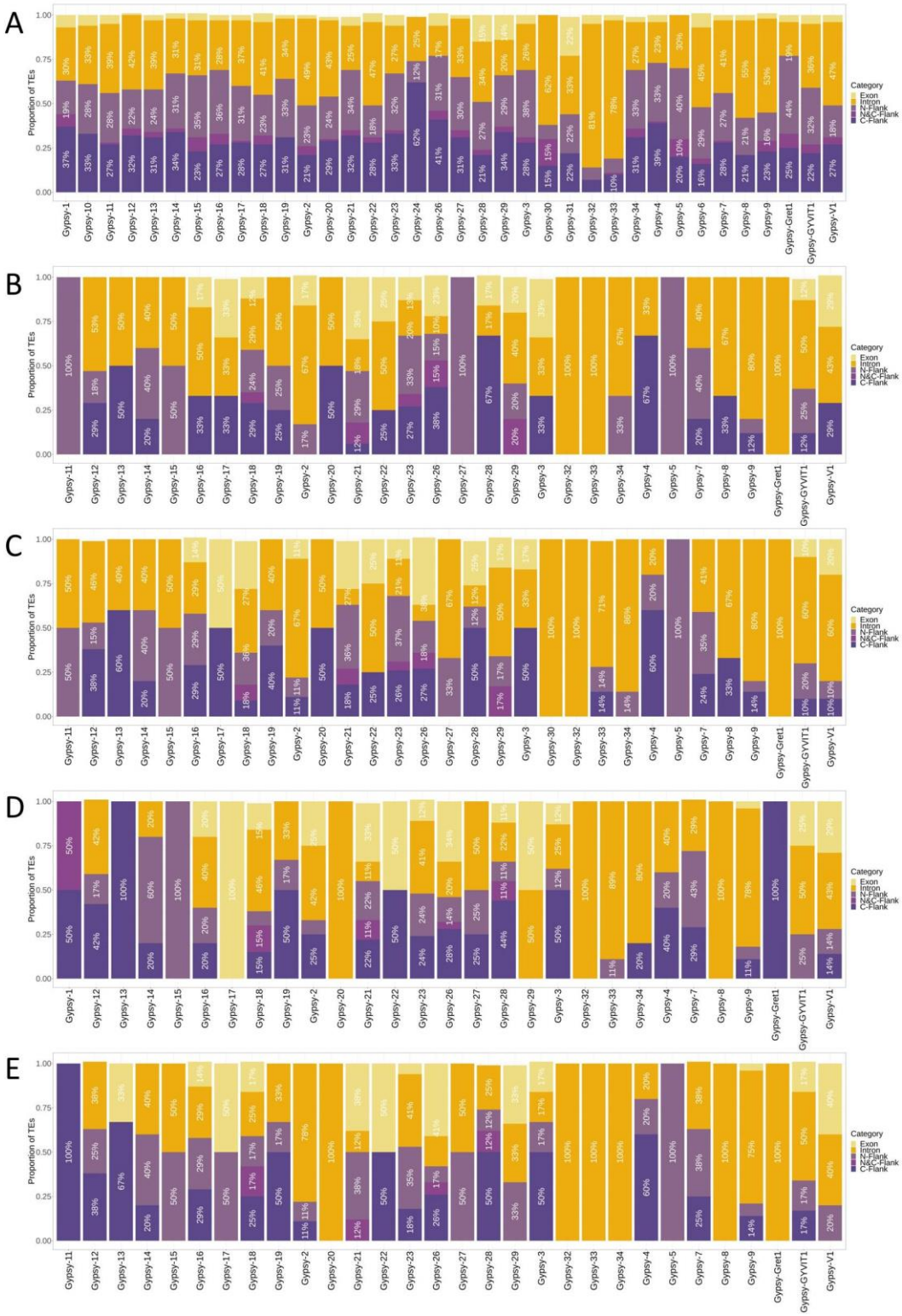



**Figure C.11 Genic and intergenic distribution of annotated TEs and expression candidates of Gypsy**

(A) Location distribution of all annotated genic Gypsy in the reference genome. (B-E) Location distribution of genic Gypsy expression candidates of T=0 (B), mock (C), yeast (D), and Botrytis (E).

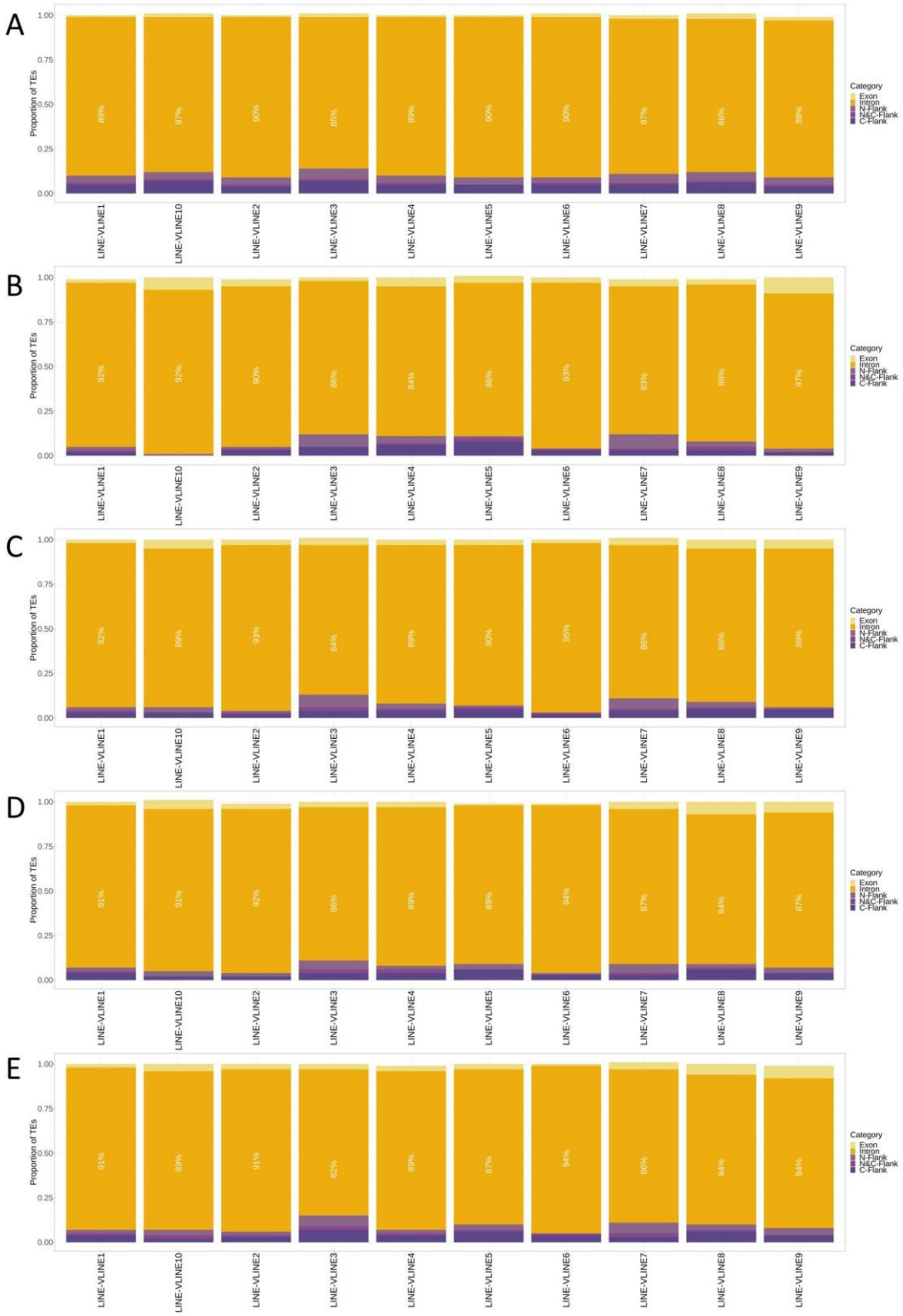



**Figure C.12 Genic and intergenic distribution of annotated TEs and expression candidates of LINE**

(A) Location distribution of all annotated genic LINE in the reference genome. (B-E) Location distribution of genic LINE expression candidates of T=0 (B), mock (C), yeast (D), and Botrytis (E).

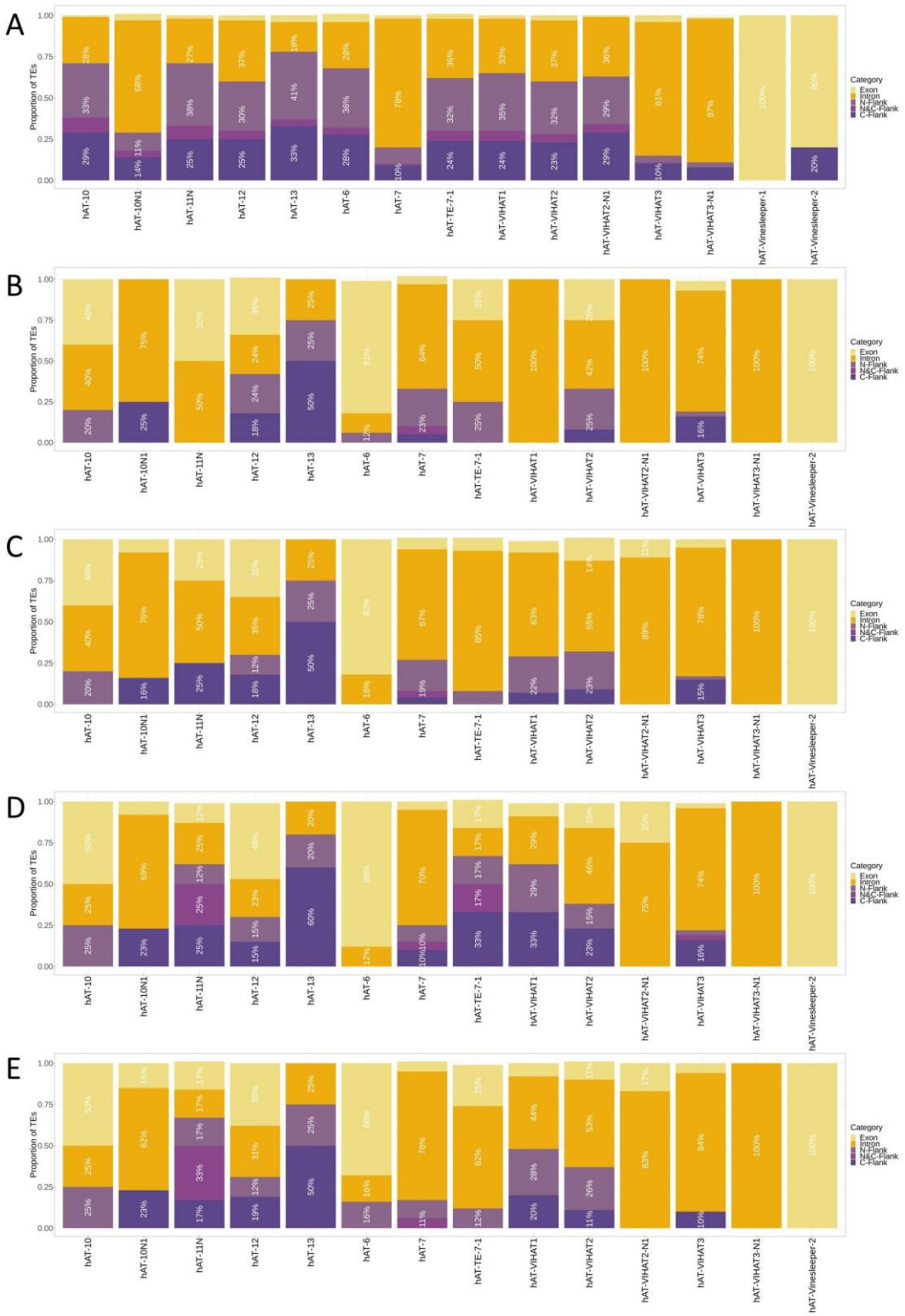



**Figure C.13 Genic and intergenic distribution of annotated TEs and expression candidates of hAT**

(A) Location distribution of all annotated genic hAT in the reference genome. (B-E) Location distribution of genic hAT expression candidates of T=0 (B), mock (C), yeast (D), and Botrytis (E).

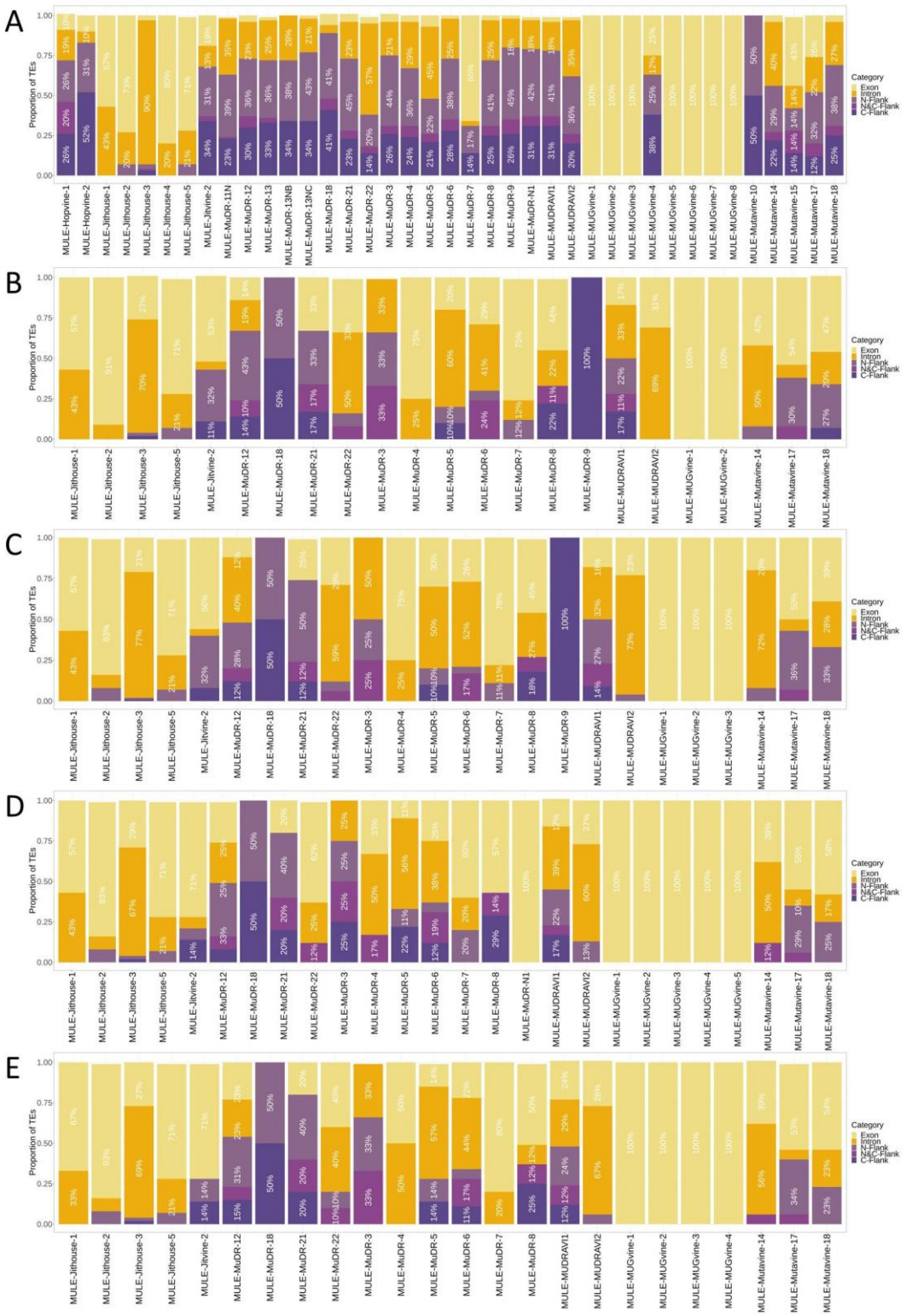



**Figure C.14 Genic and intergenic distribution of annotated TEs and expression candidates of MULE**

(A) Location distribution of all annotated genic MULE in the reference genome. (B-E) Location distribution of genic MULE expression candidates of T=0 (B), mock (C), yeast (D), and Botrytis (E).

## C.7 List of stress-related plant CREs

**Table C.10 List of stress-related plant CREs**

| Matrix_ID | Keyword | | Matrix_ID | Keyword |
|---|---|---|---|---|
| TF_motif_seq_0135 | heat | | TF_motif_seq_0342 | dehydration |
| TF_motif_seq_0012 | heat | | TF_motif_seq_0366 | dehydration |
| TF_motif_seq_0257 | heat | | TF_motif_seq_0408 | wound |
| TF_motif_seq_0036 | cold | | TF_motif_seq_0408 | elicitor |
| TF_motif_seq_0036 | dehydration | | TF_motif_seq_0032 | abiotic |
| TF_motif_seq_0322 | cold | | TF_motif_seq_0042 | dehydration |
| TF_motif_seq_0324 | cold | | TF_motif_seq_0106 | abiotic |
| TF_motif_seq_0324 | dehydration | | TF_motif_seq_0204 | pathogen |
| TF_motif_seq_0496 | cold | | TF_motif_seq_0298 | dehydration |
| TF_motif_seq_0496 | dehydration | | TF_motif_seq_0305 | dehydration |
| TF_motif_seq_0496 | high-salt | | TF_motif_seq_0313 | dehydration |
| TF_motif_seq_0258 | cold | | TF_motif_seq_0313 | high-salt |
| TF_motif_seq_0258 | dehydration | | TF_motif_seq_0326 | dehydration |
| TF_motif_seq_0302 | cold | | TF_motif_seq_0326 | high-salt |
| TF_motif_seq_0302 | dehydration | | TF_motif_seq_0356 | pathogen |
| TF_motif_seq_0349 | cold | | TF_motif_seq_0420 | abiotic |
| TF_motif_seq_0003 | wound | | TF_motif_seq_0300 | pathogen |
| TF_motif_seq_0006 | wound | | TF_motif_seq_0328 | low sugar |
| TF_motif_seq_0122 | wound | | TF_motif_seq_0434 | phosphate starvation |
| TF_motif_seq_0122 | tissue culture | | TF_motif_seq_0192 | sugar responsiveness |
| TF_motif_seq_0122 | elicitor | | TF_motif_seq_0218 | sugar responsiveness |
| TF_motif_seq_0148 | wound | | TF_motif_seq_0231 | sugar responsiveness |
| TF_motif_seq_0148 | elicitor | | TF_motif_seq_0272 | sugar responsiveness |
| TF_motif_seq_0195 | wound | | TF_motif_seq_0309 | low sugar |
| TF_motif_seq_0197 | wound | | TF_motif_seq_0387 | low sugar |
| TF_motif_seq_0273 | wound | | TF_motif_seq_0393 | sugar responsiveness |
| TF_motif_seq_0279 | wound | | TF_motif_seq_0417 | sugar responsiveness |
| TF_motif_seq_0401 | pathogen | | TF_motif_seq_0327 | low sugar |
| TF_motif_seq_0080 | pathogen | | TF_motif_seq_0376 | low sugar |
| TF_motif_seq_0220 | pathogen | | TF_motif_seq_0029 | UV |
| TF_motif_seq_0227 | pathogen | | TF_motif_seq_0114 | UV |
| TF_motif_seq_0275 | pathogen | | TF_motif_seq_0351 | UV |
| TF_motif_seq_0271 | pathogen | | TF_motif_seq_0413 | UV |
| TF_motif_seq_0009 | pathogen | | TF_motif_seq_0428 | UV |
| TF_motif_seq_0052 | pathogen | | TF_motif_seq_0129 | UV |
| TF_motif_seq_0116 | elicitor | | TF_motif_seq_0129 | elicitor |
| TF_motif_seq_0226 | oxidative stress | | TF_motif_seq_0172 | UV |
| TF_motif_seq_0314 | pathogen | | TF_motif_seq_0172 | elicitor |
| TF_motif_seq_0284 | dehydration | | TF_motif_seq_0111 | elicitor |
| TF_motif_seq_0350 | dehydration | | TF_motif_seq_0126 | elicitor |
| TF_motif_seq_0496 | dehydration | | TF_motif_seq_0193 | elicitor |
| TF_motif_seq_0496 | cold | | TF_motif_seq_0213 | elicitor |
| TF_motif_seq_0496 | high-salt | | TF_motif_seq_0351 | elicitor |
| TF_motif_seq_0021 | abiotic | | TF_motif_seq_0408 | elicitor |
| TF_motif_seq_0044 | pathogen | | TF_motif_seq_0091 | UV |
| TF_motif_seq_0096 | abiotic | | TF_motif_seq_0091 | elicitor |
| TF_motif_seq_0097 | abiotic | | TF_motif_seq_0142 | elicitor |
| TF_motif_seq_0133 | dehydration | | TF_motif_seq_0300 | elicitor |
| TF_motif_seq_0162 | pathogen | | TF_motif_seq_0307 | elicitor |
| TF_motif_seq_0341 | dehydration | | TF_motif_seq_0338 | elicitor |

## C.8 List of *Arabidopsis* genes used in search of epigenetic-related grapevine gene

**Table C.11 List of *Arabidopsis* genes used in search of epigenetic-related grapevine gene**

| Category | Gene name in *Arabidopsis* |
|---|---|
| RNA dependent RNA Polymerase | RDR1 |
| | RDR2 |
| | RDR4 |
| | RDR6 |
| Dicer | DCL1 |
| | DCL2 |
| | DCL3 |
| | DCL4 |
| | DCL5 |
| Argonaute | AGO1 |
| | AGO2 |
| | AGO4 |
| | AGO6 |
| | AGO10 |
| Chromatin Remodeler | DDM1 |
| DNA methyltransferase | MET1 |
| | CMT2 |
| | CMT3 |
| | DRM2 |
| DNA glycosylases | ROS1/DML1 |
| | DML2 |
| RNA Polymerase IV | NRPD1 |
| | NRPD2 |
| RNA Polymerase V | NRPE1 |
| | NRPE2 |
| GW/WG protein | NERD |
| | SPT5L |
| Other genes involving in RdDM | IDN2 |
| | CLSY1 |
| | DRD1 |
| | DMS3 |
| Histone methyltransferase | SUVH2 |
| | SUVH4 |
| | SUVH5 |
| | SUVH6 |
| | SUVH9 |

## C.9 List of grapevine gene potentially involving in epigenetic regulation

### Table C.12 List of grapevine gene potentially involving in epigenetic regulation

This table is extracted from Díaz-Riquelme et al. (2016) with the criteria described in 7.3.3.

| ID | CRIBI.Description | TAIR.Match | TAIR.Symbol | TAIR.Description |
|---|---|---|---|---|
| VIT_200s0131g00430 | SPT5 (Suppressor of Ty insertion 5) | AT5G04290 | KTF1, SPT5L | Encodes SPT5-Like, a member of the nuclear SPT5 (Suppressor of Ty insertion 5) RNA polymerase (RNAP) elongation factor family that is characterized by the presence of a carboxy-terminal extension with more than 40 WG/GW motifs. Interacts with AGO4. Required for RNA-directed DNA methylation. |
| VIT_200s0188g00085 | transcriptional activator demeter-like | AT2G36490 | DML1, ROS1 | A repressor of transcriptional gene silencing. Functions by demethylating the target promoter DNA. Interacts physically with RPA2/ROR1. In the ros1 mutants, an increase in methylation is observed in a number of gene promoters. Among the loci affected by ros1, a few (RD29A and At1g76930) are affected in cytosine methylation in all sequence contexts (CpG, CpNpG or CpNpN), although many others are affected primarily in non-CpG contexts. |
| VIT_200s0193g00085 | transcriptional activator demeter-like | AT2G36490 | DML1, ROS1 | A repressor of transcriptional gene silencing. Functions by demethylating the target promoter DNA. Interacts physically with RPA2/ROR1. In the ros1 mutants, an increase in methylation is observed in a number of gene promoters. Among the loci affected by ros1, a few (RD29A and At1g76930) are affected in cytosine methylation in all sequence contexts (CpG, CpNpG or CpNpN), although many others are affected primarily in non-CpG contexts. |
| VIT_200s0193g00165 | transcriptional activator demeter-like | AT2G36490 | DML1, ROS1 | A repressor of transcriptional gene silencing. Functions by demethylating the target promoter DNA. Interacts physically with RPA2/ROR1. In the ros1 mutants, an increase in methylation is observed in a number of gene promoters. Among the loci affected by ros1, a few (RD29A and At1g76930) are affected in cytosine methylation in all sequence contexts (CpG, CpNpG or CpNpN), although many others are affected primarily in non-CpG contexts. |
| VIT_200s0269g00135 | transcriptional activator demeter-like | AT2G36490 | DML1, ROS1 | A repressor of transcriptional gene silencing. Functions by demethylating the target promoter DNA. Interacts physically with RPA2/ROR1. In the ros1 mutants, an increase in methylation is observed in a number of gene promoters. Among the loci affected by ros1, a few (RD29A and At1g76930) are affected in cytosine methylation in all sequence contexts (CpG, CpNpG or CpNpN), although many others are affected primarily in non-CpG contexts. |
| VIT_200s0287g00010 | histone-lysine n- h3 lysine-9 specific suvh5 | AT2G35160 | SGD9, SUVH5 | Encodes SU(var)3-9 homologue 5 (SUVH5). SUVH5 has histone methyltransferase (MTase) activity in vitro and contributes to the maintenance of H3 mK9 (methylation of histone H3 at Lys-9) and CMT3-mediated non-CG methylation in vivo. This is a member of a subfamily of SET proteins that shares a conserved SRA domain. |
| VIT_200s0309g00020 | histone-lysine n- h3 lysine-9 specific suvh6 | AT2G35160 | SGD9, SUVH5 | Encodes SU(var)3-9 homologue 5 (SUVH5). SUVH5 has histone methyltransferase (MTase) activity in vitro and contributes to the maintenance of H3 mK9 (methylation of histone H3 at Lys-9) and CMT3-mediated non-CG methylation in vivo. This is a member of a subfamily of SET proteins that shares a conserved SRA domain. |
| VIT_200s0391g00040 | f-box protein fbw2 | AT4G08980 | FBW2 | Encodes an F-box gene that is a novel negative regulator of AGO1 protein levels and may play a role in ABA signalling and/or response. |
| VIT_201s0010g01240 | protein argonaute 1-like | AT1G48410 | AGO1, AtAGO1, ICU9 | Encodes an RNA Slicer that selectively recruits microRNAs and siRNAs. There is currently no evidence that AGO1 Slicer is in a high molecular weight RNA-induced silencing complex (RISC). Mutants are defective in post-transcriptional gene silencing and have pleiotropic developmental and morphological defects. Through its action on the regulation of ARF17 expression, the protein regulates genes involved at the cross talk between auxin and light signaling during adventitious root development. AGO1 seems to be targeted for degradation by silencing suppressor F-box-containing proteins from Turnip yellow virus and Cucurbit aphid-borne yellow virus. |
| VIT_201s0011g05870 | rna-dependent rna polymerase | AT1G14790 | ATRDRP1, RDR1 | Encodes RNA-dependent RNA polymerase. While not required for virus-induced post-transcriptional gene silencing (PTGS), it can promote turnover of viral RNAs in infected plants. Nomenclature according to Xie, et al. (2004). Involved in the production of Cucumber Mosaic Virus siRNAs. |
| VIT_201s0011g05880 | rna-dependent rna polymerase | AT1G14790 | ATRDRP1, RDR1 | Encodes RNA-dependent RNA polymerase. While not required for virus-induced post-transcriptional gene silencing (PTGS), it can promote turnover of viral RNAs in infected plants. Nomenclature according to Xie, et al. (2004). Involved in the production of Cucumber Mosaic Virus siRNAs. |
| VIT_201s0113g00500 | protein argonaute 7-like | AT1G69440 | AGO7, ZIP | Encodes ARGONAUTE7, a member of the ARGONAUTE family, characterised by the presence of PAZ and PIWI domains. Involved in the regulation of developmental timing. Required for the accumulation of TAS3 ta-siRNAs but not for accumulation of miR171, miR173, miR390 or mi391. Localized in mature rosette leaves and floral buds. |
| VIT_201s0150g00070 | histone-lysine n- h3 lysine-9 specific suvh4-like | AT5G13960 | KYP, SDG33, SUVH4 | Encodes a histone 3 lysine 9 specific methyltransferase involved in the maintenance of DNA methylation. SUVH4/KYP is a SU(VAR)3-9 homolog, a SET domain protein. Known SET domain proteins are involved in epigenetic control of gene expression. There are 10 SUVH genes in Arabidopsis and members of this subfamily of the SET proteins have an additional conserved SRA domain. In kyp mutants, there is a loss of CpNpG methylation. The protein was shown to bind to methylated cytosines of CG, CNG and CNN motifs via its SRA domain but has a preference for the latter two. There is also evidence that KYP/SUVH4 might be involved in the telomerase-independent process known as Alternative Lengthening of Telomeres. |

| | | | | |
|---|---|---|---|---|
| VIT_202s0025g04525 | nuclear rna polymerase d1a | AT1G63020 | NRPD1, NRPD1A, POL, SDE4, SMD2 | Encodes one of two alternative largest subunits of a putative plant-specific RNA polymerase IV (aka RNA polymerase D). Required for posttranscriptional gene silencing. |
| VIT_202s0025g04530 | nuclear rna polymerase d1a | AT1G63020 | NRPD1, NRPD1A, POL, SDE4, SMD2 | Encodes one of two alternative largest subunits of a putative plant-specific RNA polymerase IV (aka RNA polymerase D). Required for posttranscriptional gene silencing. |
| VIT_202s0033g00610 | chromomethylase 2 | AT4G19020 | CMT2 | NA |
| VIT_203s0038g00610 | zinc finger ccch domain-containing protein 44-like | AT2G16485 | NERD | NA |
| VIT_203s0038g00830 | protein argonaute 7-like | AT1G69440 | AGO7, ZIP | Encodes ARGONAUTE7, a member of the ARGONAUTE family, characterised by the presence of PAZ and PIWI domains. Involved in the regulation of developmental timing. Required for the accumulation of TAS3 ta-siRNAs but not for accumulation of miR171, miR173, miR390 or mi391. Localized in mature rosette leaves and floral buds. |
| VIT_203s0038g04240 | histone deacetylase | AT4G38130 | ATHD1, ATHDA19, HD1, HDA1, HDA19, RPD3A | Encodes a histone deacetylase that enhances AtERF7-mediated transcriptional repression. Binds SIM3 and ERF7. Expressed in the nucleus in most tissues examined and throughout the life of the plant. Involved in jasmonic acid and ethylene dependent pathogen resistance. The sequence in GenBank has 17 AG dinucleotide repeats missing, which is also missing in Ler shotgun sequence from Cereon. Although it is annotated to be in Columbia, the GB sequence is probably not of Columbia origin. Plays a role in embryogenesis as mutants grown at higher temperatures display abnormalities in the organization of the root and shoot. Plant lines expressing an RNAi construct targeted against HDA19 shows some resistance to agrobacterium-mediated root transformation. |
| VIT_203s0038g04290 | dna repair and recombination protein rad54-like | AT2G16390 | CHR35, DMS1, DRD1 | Putative chromatin remodeling protein, member of a plant-specific subfamily of SWI2/SNF2-like proteins. Mutations nearly eliminate non-CpG methylation at a target promoter but do not affect rDNA or centromere methylation. Cooperates with PolIVb to facilitate RNA-directed de novo methylation and silencing of homologous DNA. Endogenous targets include intergenic regions near retrotransposon LTRs or short RNA encoding sequences that might epigenetically regulate adjacent genes. May be used to establish a basal yet reversible level of silencing in euchromatin. |
| VIT_203s0132g00020 | histone deacetylase hdt1-like | AT5G03740 | HD2C, HDT3 | HD2-type histone deacetylase HDAC. Involved in the ABA and stress responses. Mediates transcriptional repression |
| VIT_204s0008g00910 | histone deacetylase 2 | AT5G26040 | HDA2 | Class III RPD3 type protein. Encodes HDA2, a member of the histone deacetylase family proteins. |
| VIT_204s0008g02150 | protein dicer-like 3 | AT3G43920 | ATDCL3, DCL3 | Encodes a ribonuclease III family protein that is required for endogenous RDR2-dependent siRNA (but not miRNA) formation. |
| VIT_204s0008g05430 | rna-dependent rna polymerase 6 | AT3G49500 | RDR6, SDE1, SGS2 | Encodes RNA-dependent RNA polymerase. Involved in trans-acting siRNA and other siRNA biogenesis. Required for post-transcriptional gene silencing and natural virus resistance. |
| VIT_204s0008g06580 | dna-directed rna polymerase d subunit 2a-like | AT3G23780 | DMS2, DRD2, NRPD2, NRPD2A, NRPE2, OCP1 | This gene encodes the second largest, catalytic subunit of the nuclear DNA-dependent RNA polymerase IV (aka RNA polymerase D). The NRPD2 protein is found at nuclear foci that overlap or are adjacent to chromocentromeres but are not fully coincident with chromocentromeres. The loss of NRPD2 leads to the loss of cytosine methylation at pericentromeric 5S genes and AtSN1 retroelements but has no discernible effect on centromere repeat methylation. This suggests that Pol IV primarily affects facultative heterochromatin rather than constitutive heterochromatin. |
| VIT_204s0023g00920 | protein dicer-like 2 | AT3G03300 | ATDCL2, DCL2 | Encodes a Dicer-like protein that functions in the antiviral silencing response in turnip-crinkle virus-infected plants but not in TMV or CMV-strain-Y-infected plants. Involved in the production of ta-siRNAs. Partially antagonizes the production of miRNAs by DCL1. Substitutes for DCL4 to produce viral siRNA when DCL4 is missing or inhibited. Able to produce siRNAs but not miRNAs. |
| VIT_204s0023g01610 | atp-dependent dna helicase ddm1-like | AT5G66750 | ATDDM1, CHA1, CHR01, CHR1, DDM1, SOM1, SOM4 | Protein is similar to SWI2/SNF2 chromatin remodeling proteins. DDM1 is appears to act as a chromatin-remodeling ATPase involved in cytosine methylation in CG and non-CG contexts. Involved in gene silencing and maintenance of DNA methylation and histone methylation. Hypomethylation of many genomic regions occurs in ddm1 mutants, and can cause several phenotypic abnormalities, but some loci, such as BONSAI (At1g73177) can be hypermethylated in ddm1 mutants after several generations, leading to different phenotypes. DDM1 might be involved in establishing a heterochromain boundary. A line expressing an RNAi targeted against DDM1 shows some resistance to agrobacterium-mediated root transformation. |
| VIT_204s0023g02950 | zinc finger ccch domain-containing protein 44-like | AT2G16485 | NERD | NA |
| VIT_204s0044g01510 | histone deacetylase | AT4G33470 | ATHDA14, HDA14 | Encodes HDA14, a member of the histone deacetylase family proteins. |
| VIT_205s0020g03840 | NA | AT3G22680 | RDM1 | Encodes RNA-DIRECTED DNA METHYLATION 1 (RDM1), forming a complex with DMS3 (AT3G49250) and DRD1 (AT2G16390). This complex is termed DDR. The DDR complex is required for polymerase V transcripts and RNA-directed DNA methylation. |
| VIT_205s0020g04190 | protein argonaute 10-like | AT5G43810 | AGO10, PNH, ZLL | Encodes a member of the EIF2C (elongation initiation factor 2c)/ Argonaute class of proteins. Required to establish the central-peripheral organization of the embryo apex. Along with WUS and CLV genes, controls the relative organization of central zone and peripheral zone cells in meristems. Acts in |

| | | | | |
|---|---|---|---|---|
| | | | | embryonic provascular tissue potentiating WUSCHEL function during meristem development in the embryo. |
| VIT_205s0020g04760 | rna polymerase rpb7 n-terminal domain-containing protein | AT4G14660 | NRPE7 | Non-catalytic subunit specific to DNA-directed RNA polymerase V; homologous to budding yeast RPB7 |
| VIT_205s0049g02220 | histone-lysine n- h3 lysine-9 specific suvh9 | AT2G33290 | ATSUVH2, SDG3, SUVH2 | Encodes a SU(VAR)3-9 homolog, a SET domain protein. Known SET domain proteins are involved in epigenetic control of gene expression and act as histone methyltransferases. There are 10 SUVH genes in Arabidopsis and members of this subfamily of the SET proteins have an additional conserved SRA domain. Gene is expressed in rosettes, stems, floral buds, and flowers by RT-PCR. |
| VIT_206s0004g01080 | dna (cytosine-5)-methyltransferase | AT1G69770 | CMT3 | Encodes a chromomethylase involved in methylating cytosine residues at non-CG sites. Involved in preferentially methylating transposon-related sequences, reducing their mobility. CMT3 interacts with an Arabidopsis homologue of HP1 (heterochromatin protein 1), which in turn interacts with methylated histones. Involved in gene silencing. |
| VIT_206s0004g08480 | dna repair and recombination protein rad54-like | AT2G16390 | CHR35, DMS1, DRD1 | Putative chromatin remodeling protein, member of a plant-specific subfamily of SWI2/SNF2-like proteins. Mutations nearly eliminate non-CpG methylation at a target promoter but do not affect rDNA or centromere methylation. Cooperates with PolIVb to facilitate RNA-directed de novo methylation and silencing of homologous DNA. Endogenous targets include intergenic regions near retrotransposon LTRs or short RNA encoding sequences that might epigenetically regulate adjacent genes. May be used to establish a basal yet reversible level of silencing in euchromatin. |
| VIT_206s0009g01200 | protein argonaute 4 | AT2G27040 | AGO4, OCP11 | AGO4 is a member of a class of PAZ/PIWI domain containing proteins involved in siRNA mediated gene silencing.Loss of function mutations have reduced site specific CpNpG and CpHpH methylation and increased susceptibility to bacterial pathogens. |
| VIT_206s0061g01040 | eukaryotic translation initiation factor | AT2G27880 | AGO5, AtAGO5 | NA |
| VIT_206s0061g01240 | histone deacetylase | AT3G44750 | ATHD2A, HD2A, HDA3, HDT1 | Encodes a histone deacetylase. Controls the development of adaxial/abaxial leaf polarity. Two lines with RNAi-directed against this gene show reduced Agrobacterium-mediated DNA transformation of the roots. |
| VIT_206s0061g01510 | histone deacetylase | AT3G44680 | HDA09, HDA9 | Class I RPD3 type protein |
| VIT_206s0080g00210 | histone deacetylase | AT1G08460 | ATHDA8, HDA08, HDA8 | NA |
| VIT_207s0005g01490 | histone-lysine n-methyltransferase eza1-like | AT4G02020 | EZA1, SDG10, SWN | Encodes a polycomb group protein. Forms part of a large protein complex that can include VRN2 (VERNALIZATION 2), VIN3 (VERNALIZATION INSENSITIVE 3) and polycomb group proteins FERTILIZATION INDEPENDENT ENDOSPERM (FIE) and CURLY LEAF (CLF). The complex has a role in establishing FLC (FLOWERING LOCUS C) repression during vernalization. Performs a partially redundant role to MEA in controlling seed initiation by helping to suppress central cell nucleus\002endosperm proliferation within the FG. |
| VIT_207s0031g02510 | sirtuin 2 | AT5G09230 | AtSRT2, SRT2 | Encodes SRT2, a member of the SIR2 (sirtuin) family HDAC (histone deacetylase) (SRT1/AT5g55760, SRT2/AT5G09230). |
| VIT_207s0130g00190 | protein suppressor of gene silencing 3-like | AT5G23570 | ATSGS3, SGS3 | Required for posttranscriptional gene silencing and natural virus resistance.SGS3 is a member of an 'unknown' protein family. Members of this family have predicted coiled coiled domains suggesting oligomerization and a potential zinc finger domain. Involved in the production of trans-acting siRNAs, through direct or indirect stabilization of cleavage fragments of the primary ta-siRNA transcript. Acts before RDR6 in this pathway. |
| VIT_207s0130g00380 | dna (cytosine-5)-methyltransferase | AT5G49160 | DDM2, DMT01, DMT1, MET1, MET2, METI | Encodes a cytosine methyltransferase MET1. Required for silencing of FWA paternal allele in endosperm. Two lines with RNAi constructs directed against DMT1 have reduced agrobacterium-mediated tumor formation. |
| VIT_207s0130g00390 | dna (cytosine-5)-methyltransferase | AT5G49160 | DDM2, DMT01, DMT1, MET1, MET2, METI | Encodes a cytosine methyltransferase MET1. Required for silencing of FWA paternal allele in endosperm. Two lines with RNAi constructs directed against DMT1 have reduced agrobacterium-mediated tumor formation. |
| VIT_208s0007g04360 | argonaute protein group | AT2G27880 | AGO5, AtAGO5 | NA |
| VIT_208s0007g05540 | eukaryotic rpb5 rna polymerase subunit family protein | AT3G57080 | NRPE5, RPB5B | Non-catalytic subunit unique to Nuclear DNA-dependent RNA polymerase V; homologous to budding yeast RPB5. |
| VIT_208s0007g06800 | dna (cytosine-5)-methyltransferase | AT1G69770 | CMT3 | Encodes a chromomethylase involved in methylating cytosine residues at non-CG sites. Involved in preferentially methylating transposon-related sequences, reducing their mobility. CMT3 interacts with an Arabidopsis homologue of HP1 (heterochromatin protein 1), which in turn interacts with methylated histones. Involved in gene silencing. |
| VIT_208s0040g00070 | protein argonaute 4 | AT2G27040 | AGO4, OCP11 | AGO4 is a member of a class of PAZ/PIWI domain containing proteins involved in siRNA mediated gene silencing.Loss of function mutations have reduced site specific CpNpG and CpHpH methylation and increased susceptibility to bacterial pathogens. |
| VIT_208s0056g01550 | NA | AT3G62800 | ATTIF3K1, DRB4 | Encodes a nuclear dsRNA-binding protein that interacts specifically with DCL4. May regulate DCL4 function and thereby affect miRNA biogenesis. DRB4 interacts with the P6 viral protein from Cauliflower mosaic virus and may be a target of viral silencing suppression. |
| VIT_209s0018g00126 | transcriptional activator demeter- | AT2G36490 | DML1, ROS1 | A repressor of transcriptional gene silencing. Functions by demethylating the target promoter DNA. Interacts physically with |

| | | | | |
|---|---|---|---|---|
| | like | | | RPA2/ROR1. In the ros1 mutants, an increase in methylation is observed in a number of gene promoters. Among the loci affected by ros1, a few (RD29A and At1g76930) are affected in cytosine methylation in all sequence contexts (CpG, CpNpG or CpNpN), although many others are affected primarily in non-CpG contexts. |
| VIT_209s0018g00158 | transcriptional activator demeter-like | AT2G36490 | DML1, ROS1 | A repressor of transcriptional gene silencing. Functions by demethylating the target promoter DNA. Interacts physically with RPA2/ROR1. In the ros1 mutants, an increase in methylation is observed in a number of gene promoters. Among the loci affected by ros1, a few (RD29A and At1g76930) are affected in cytosine methylation in all sequence contexts (CpG, CpNpG or CpNpN), although many others are affected primarily in non-CpG contexts. |
| VIT_210s0003g04746 | dna repair and recombination protein rad54-like | AT2G16390 | CHR35, DMS1, DRD1 | Putative chromatin remodeling protein, member of a plant-specific subfamily of SWI2/SNF2-like proteins. Mutations nearly eliminate non-CpG methylation at a target promoter but do not affect rDNA or centromere methylation. Cooperates with PolIVb to facilitate RNA-directed de novo methylation and silencing of homologous DNA. Endogenous targets include intergenic regions near retrotransposon LTRs or short RNA encoding sequences that might epigenetically regulate adjacent genes. May be used to establish a basal yet reversible level of silencing in euchromatin. |
| VIT_210s0003g04813 | dna repair and recombination protein rad54-like | AT2G16390 | CHR35, DMS1, DRD1 | Putative chromatin remodeling protein, member of a plant-specific subfamily of SWI2/SNF2-like proteins. Mutations nearly eliminate non-CpG methylation at a target promoter but do not affect rDNA or centromere methylation. Cooperates with PolIVb to facilitate RNA-directed de novo methylation and silencing of homologous DNA. Endogenous targets include intergenic regions near retrotransposon LTRs or short RNA encoding sequences that might epigenetically regulate adjacent genes. May be used to establish a basal yet reversible level of silencing in euchromatin. |
| VIT_210s0042g01200 | protein argonaute 2-like | AT1G31280 | AGO2, AtAGO2 | An Argonaute gene |
| VIT_210s0168g00007 | protein argonaute 7-like | AT1G69440 | AGO7, ZIP | Encodes ARGONAUTE7, a member of the ARGONAUTE family, characterised by the presence of PAZ and PIWI domains. Involved in the regulation of developmental timing. Required for the accumulation of TAS3 ta-siRNAs but not for accumulation of miR171, miR173, miR390 or mi391. Localized in mature rosette leaves and floral buds. |
| VIT_211s0016g04620 | protein argonaute pnh1-like | AT5G43810 | AGO10, PNH, ZLL | Encodes a member of the EIF2C (elongation initiation factor 2c)/ Argonaute class of proteins. Required to establish the central-peripheral organization of the embryo apex. Along with WUS and CLV genes, controls the relative organization of central zone and peripheral zone cells in meristems. Acts in embryonic provascular tissue potentiating WUSCHEL function during meristem development in the embryo. |
| VIT_211s0149g00100 | dicer-like protein 4 | AT5G20320 | ATDCL4, DCL4 | Encodes an RNase III-like enzyme that catalyzes processing of trans-acting small interfering RNA precursors in a distinct small RNA biogenesis pathway. The protein is also involved in the production of 21-nt primary siRNAs from both inverted-repeat constructs and endogenous sequences, as well as the RDR6-dependent 21-nt secondary siRNAs involved in long-range cell-to-cell signaling. It binds DRB4, a ds-RNA binding protein. |
| VIT_211s0149g00120 | dicer-like protein 4 | AT5G20320 | ATDCL4, DCL4 | Encodes an RNase III-like enzyme that catalyzes processing of trans-acting small interfering RNA precursors in a distinct small RNA biogenesis pathway. The protein is also involved in the production of 21-nt primary siRNAs from both inverted-repeat constructs and endogenous sequences, as well as the RDR6-dependent 21-nt secondary siRNAs involved in long-range cell-to-cell signaling. It binds DRB4, a ds-RNA binding protein. |
| VIT_212s0034g02560 | dna (cytosine-5)-methyltransferase | AT5G49160 | DDM2, DMT01, DMT1, MET1, MET2, METI | Encodes a cytosine methyltransferase MET1. Required for silencing of FWA paternal allele in endosperm. Two lines with RNAi constructs directed against DMT1 have reduced agrobacterium-mediated tumor formation. |
| VIT_212s0035g00010 | dna (cytosine-5)-methyltransferase | AT5G49160 | DDM2, DMT01, DMT1, MET1, MET2, METI | Encodes a cytosine methyltransferase MET1. Required for silencing of FWA paternal allele in endosperm. Two lines with RNAi constructs directed against DMT1 have reduced agrobacterium-mediated tumor formation. |
| VIT_212s0035g01755 | dna methyltransferase | AT5G49160 | DDM2, DMT01, DMT1, MET1, MET2, METI | Encodes a cytosine methyltransferase MET1. Required for silencing of FWA paternal allele in endosperm. Two lines with RNAi constructs directed against DMT1 have reduced agrobacterium-mediated tumor formation. |
| VIT_212s0059g01430 | argonaute protein group | AT2G32940 | AGO6 | Encodes a nuclear localized 879-amino-acid protein that contains conserved PAZ and PIWI domains that is important for the accumulation of specific heterochromatin-related siRNAs, and for DNA methylation and transcriptional gene silencing. |
| VIT_212s0059g02300 | fha domain-containing protein ddl | AT3G20550 | DDL | Encodes a nuclear localized FHA (forhkead) domain containing protein.Mutant plants have shortened roots, delayed flowering time, altered floral organ number, defective floral organs and reduced fertility.Ddl mutants also show reduced levels of pri-miRNAs as well as mature miRNAs suggesting involvement in biogenesis of miRNAs. DDL does not affect transcription of miRNAs directly but may act through other proteins such as DCL. |
| VIT_213s0019g01610 | xh xs domain-containing protein | AT3G48670 | IDN2, RDM12 | Encodes IDN2 (INVOLVED IN DE NOVO 2), a double-stranded RNA-binding protein involved in de novo methylation and small interfering RNA (siRNA)-mediated maintenance methylation. IND2 is a component of the RNA-directed DNA methylation pathway. |
| VIT_213s0019g04950 | suppressor of | AT5G04290 | KTF1, SPT5L | Encodes SPT5-Like, a member of the nuclear SPT5 (Suppressor of Ty insertion 5) RNA polymerase (RNAP) elongation factor family that is characterized by the presence of a carboxy-terminal extension with more than 40 WG/GW |

| | | | | |
|---|---|---|---|---|
| | | | | motifs. Interacts with AGO4. Required for RNA-directed DNA methylation. |
| VIT_213s0047g00120 | set domain protein | AT1G73100 | SDG19, SUVH3 | Encodes a SU(VAR)3-9 homolog, a SET domain protein. Known SET domain proteins are involved in epigenetic control of gene expression and act as histone methyltransferases. There are 10 SUVH genes in Arabidopsis and members of this subfamily of the SET proteins have an additional conserved SRA domain. |
| VIT_213s0067g03120 | rna polymerase iv subunit | AT2G40030 | ATNRPD1B, DMS5, DRD3, NRPD1B, NRPE1 | Encodes the unique largest subunit of nuclear DNA-dependent RNA polymerase V; homologous to budding yeast RPB1 and the E. coli RNA polymerase beta prime subunit. Required for normal RNA-directed DNA methylation at non-CG methylation sites and transgene silencing. |
| VIT_213s0106g00170 | histone deacetylase | AT4G33470 | ATHDA14, HDA14 | Encodes HDA14, a member of the histone deacetylase family proteins. |
| VIT_213s0175g00140 | protein argonaute 4 | AT2G27040 | AGO4, OCP11 | AGO4 is a member of a class of PAZ/PIWI domain containing proteins involved in siRNA mediated gene silencing.Loss of function mutations have reduced site specific CpNpG and CpHpH methylation and increased susceptibility to bacterial pathogens. |
| VIT_214s0006g01820 | histone deacetylase | AT4G38130 | ATHD1, ATHDA19, HD1, HDA1, HDA19, RPD3A | Encodes a histone deacetylase that enhances AtERF7-mediated transcriptional repression. Binds SIM3 and ERF7. Expressed in the nucleus in most tissues examined and throughout the life of the plant. Involved in jasmonic acid and ethylene dependent pathogen resistance. The sequence in GenBank has 17 AG dinucleotide repeats missing, which is also missing in Ler shotgun sequence from Cereon. Although it is annotated to be in Columbia, the GB sequence is probably not of Columbia origin. Plays a role in embryogenesis as mutants grown at higher temperatures display abnormalities in the organization of the root and shoot. Plant lines expressing an RNAi construct targeted against HDA19 shows some resistance to agrobacterium-mediated root transformation. |
| VIT_214s0006g02120 | suppressor of | AT5G04290 | KTF1, SPT5L | Encodes SPT5-Like, a member of the nuclear SPT5 (Suppressor of Ty insertion 5) RNA polymerase (RNAP) elongation factor family that is characterized by the presence of a carboxy-terminal extension with more than 40 WG/GW motifs. Interacts with AGO4. Required for RNA-directed DNA methylation. |
| VIT_214s0030g01580 | 5 -3 exoribonuclease 4-like | AT1G54490 | AIN1, ATXRN4, EIN5, XRN4 | Involved in the ethylene response. XRN4 does not appear to regulate ethylene signaling via an RNA-INDUCED SILENCING COMPLEX-based RNA silencing mechanism but acts by independent means. Endogenous suppressor of posttranscriptional gene silencing. |
| VIT_214s0066g01040 | dna (cytosine-5)-methyltransferase drm2 | AT5G14620 | DMT7, DRM2 | A putative DNA methyltransferase with rearranged catalytic domains; similar to mammalian DNMT3 methyltransferases; contains UBA domains. The 3'-end proximal part of the gene coding region is highly methylated at both adenine and cytosine residues. |
| VIT_214s0068g01070 | histone-lysine n- h3 lysine-9 specific suvh4-like | AT5G13960 | KYP, SDG33, SUVH4 | Encodes a histone 3 lysine 9 specific methyltransferase involved in the maintenance of DNA methylation. SUVH4/KYP is a SU(VAR)3-9 homolog, a SET domain protein. Known SET domain proteins are involved in epigenetic control of gene expression. There are 10 SUVH genes in Arabidopsis and members of this subfamily of the SET proteins have an additional conserved SRA domain. In kyp mutants, there is a loss of CpNpG methylation. The protein was shown to bind to methylated cytosines of CG, CNG and CNN motifs via its SRA domain but has a preference for the latter two. There is also evidence that KYP/SUVH4 might be involved in the telomerase-independent process known as Alternative Lengthening of Telomeres. |
| VIT_215s0021g00610 | histone deacetylase | AT3G44680 | HDA09, HDA9 | Class I RPD3 type protein |
| VIT_215s0048g01290 | histone deacetylase complex subunit sap18 | AT2G45640 | ATSAP18, SAP18 | Involved in the regulation of salt stress. Expression of AtSAP18 is induced by NaCl, cold, drought, ABA, and ethylene treatment. AtSAP18 and HDA19 associate with ERF3 and ERF4 both in vitro and in vivo. |
| VIT_215s0048g02380 | endoribonuclease dicer homolog 1-like | AT1G01040 | ASU1, ATDCL1, CAF, DCL1, EMB60, EMB76, SIN1, SUS1 | Encodes a Dicer homolog. Dicer is a RNA helicase involved in microRNA processing. Mutations in this locus can result in embryo lethality. Embryo shape at seed maturity is globular-elongate. Other mutants convert the floral meristems to an indeterminate state, others yet show defects in ovule development. mRNA is expressed in all shoot tissues. DCL1 is able to produce miRNAs and siRNAs. |
| VIT_216s0013g00310 | histone-lysine n- h3 lysine-9 specific suvh5 | AT2G35160 | SGD9, SUVH5 | Encodes SU(var)3-9 homologue 5 (SUVH5). SUVH5 has histone methyltransferase (MTase) activity in vitro and contributes to the maintenance of H3 mK9 (methylation of histone H3 at Lys-9) and CMT3-mediated non-CG methylation in vivo. This is a member of a subfamily of SET proteins that shares a conserved SRA domain. |
| VIT_216s0013g01550 | multicopy suppressor of ira1 | AT5G58230 | ATMSI1, MEE70, MSI1 | Encodes a WD-40 repeat containing protein that functions in chromatin assembly as part of the CAF1 and FIE complex. Mutants exhibit parthenogenetic development that includes proliferation of unfertilized endosperm and embryos. In heterozygous plants 50% of embryos abort. Of the aborted embryos the early aborted class are homozygous and the later aborting lass are heterozygotes in which the defective allele is maternally transmitted. Other phenotypes include defects in ovule morphogenesis and organ initiation,as well as increased levels of heterochromatic DNA. MSI1 is needed for the transition to flowering. In Arabidopsis, the three CAF-1 subunits are encoded by FAS1, FAS2 and, most likely, MSI1, respectively. Mutations in FAS1 or FAS2 lead to increased frequency of homologous recombination and T-DNA integration in Arabidopsis. In the ovule, the MSI1 transcripts are accumulated at their highest level before fertilization and gradually decrease after fertilization. MSI is biallelically expressed, the paternall allele is expressed in the endosperm and embryo and is not imprinted. MSI1 forms a complex with RBR1 that is required for activation of |

| | | | | |
|---|---|---|---|---|
| | | | | the imprinted genes FIS2 and FWA. This activation is mediated by MSI1/RBR1 mediated repression of MET1. |
| VIT_216s0039g02460 | chromomethylase 2 | AT4G19020 | CMT2 | NA |
| VIT_216s0039g02640 | 5 -3 exoribonuclease 2 | AT1G54490 | AIN1, ATXRN4, EIN5, XRN4 | Involved in the ethylene response. XRN4 does not appear to regulate ethylene signaling via an RNA-INDUCED SILENCING COMPLEX-based RNA silencing mechanism but acts by independent means. Endogenous suppressor of posttranscriptional gene silencing. |
| VIT_216s0098g01820 | protein defective in meristem silencing 3 | AT3G49250 | DMS3, IDN1 | Similar to hinge-domain region of structural maintenance of chromosomes (SMC)proteins.Putative chromosome architecture protein that can potentialy link nucleic acids in facilitating an RNA1-mediated epigenetic modification involving secondary siRNA and spreading of DNA methylation. |
| VIT_217s0000g04120 | histone deacetylase 15-like | AT3G18520 | ATHDA15, HDA15 | Encodes a protein with similarity to histone deacetylases. Plants expressing RNAi directed against this gene show a moderate resistance to agrobacterium-mediated root transformation. |
| VIT_217s0000g07280 | histone deacetylase | AT5G61060 | ATHDA5, HDA05, HDA5 | Encodes a member of the histone deacetylase family. |
| VIT_217s0000g09070 | histone deacetylase | AT5G63110 | ATHDA6, AXE1, HDA6, RPD3B, RTS1, SIL1 | RPD3-like histone deacetylase. HDA6 mutations specifically increase the expression of auxin-responsive transgenes, suggesting a role in transgene silencing. |
| VIT_217s0053g00680 | protein argonaute | AT1G48410 | AGO1, AtAGO1, ICU9 | Encodes an RNA Slicer that selectively recruits microRNAs and siRNAs. There is currently no evidence that AGO1 Slicer is in a high molecular weight RNA-induced silencing complex (RISC). Mutants are defective in post-transcriptional gene silencing and have pleiotropic developmental and morphological defects. Through its action on the regulation of ARF17 expression, the protein regulates genes involved at the cross talk between auxin and light signaling during adventitious root development. AGO1 seems to be targeted for degradation by silencing suppressor F-box-containing proteins from Turnip yellow virus and Cucurbit aphid-borne yellow virus. |
| VIT_218s0001g06220 | f-box protein fbw2 | AT4G08980 | FBW2 | Encodes an F-box gene that is a novel negative regulator of AGO1 protein levels and may play a role in ABA signalling and/or response. |
| VIT_219s0014g01840 | protein argonaute 1-like | AT1G48410 | AGO1, AtAGO1, ICU9 | Encodes an RNA Slicer that selectively recruits microRNAs and siRNAs. There is currently no evidence that AGO1 Slicer is in a high molecular weight RNA-induced silencing complex (RISC). Mutants are defective in post-transcriptional gene silencing and have pleiotropic developmental and morphological defects. Through its action on the regulation of ARF17 expression, the protein regulates genes involved at the cross talk between auxin and light signaling during adventitious root development. AGO1 seems to be targeted for degradation by silencing suppressor F-box-containing proteins from Turnip yellow virus and Cucurbit aphid-borne yellow virus. |
| VIT_219s0015g00570 | sirtuin 1 | AT5G55760 | SRT1 | Encodes SRT1, a member of the SIR2 (sirtuin) family HDAC (histone deacetylase) (SRT1/AT5g55760, SRT2/AT5G09230). |
| VIT_219s0093g00140 | suppressor of | AT5G04290 | KTF1, SPT5L | Encodes SPT5-Like, a member of the nuclear SPT5 (Suppressor of Ty insertion 5) RNA polymerase (RNAP) elongation factor family that is characterized by the presence of a carboxy-terminal extension with more than 40 WG/GW motifs. Interacts with AGO4. Required for RNA-directed DNA methylation. |

## C.10 Differential expression analysis of mock and 4PBA expression candidates

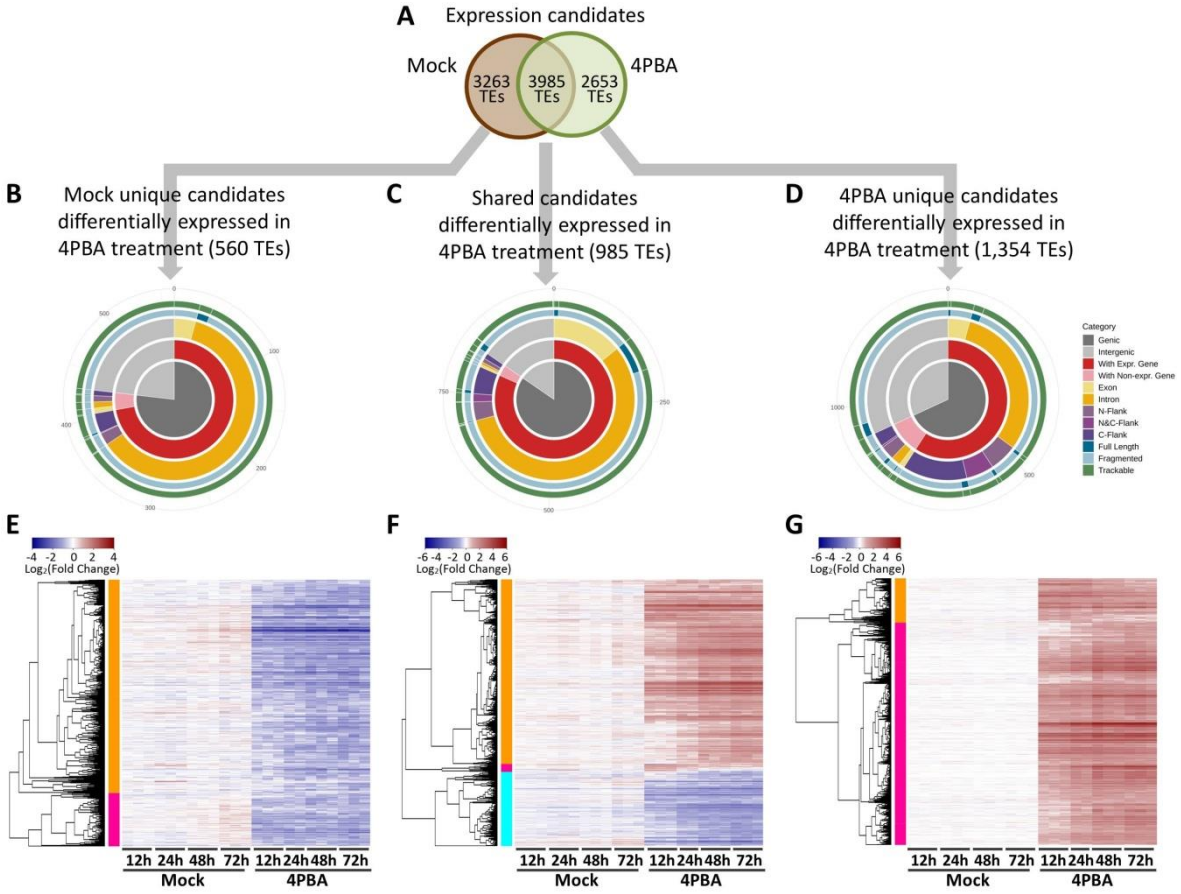



**Figure C.15 Location and expression pattern of DETEs of mock and 4PBA treatment**

**(A)** Mock and 4PBA expression candidates were sorted into three groups: mock-unique, shared, and 4PBA-unique. **(B-D)** For each group, DETEs were extracted and categorized by location, the transcriptional activity of co-localized genes, presence/absence of unique-mapping reads, as well as TE integrity. The number of DETEs in each group was indicated. **(E-F)** Heatmaps of DETEs demonstrate the logarithmically transformed fold change comparing to T=0. Therefore, the white colour indicates the expression level as same as at T=0.

## C.11 Comparisons of gene and TE expression quantified from ONT versus from Illumina Truseq sequencing libraries

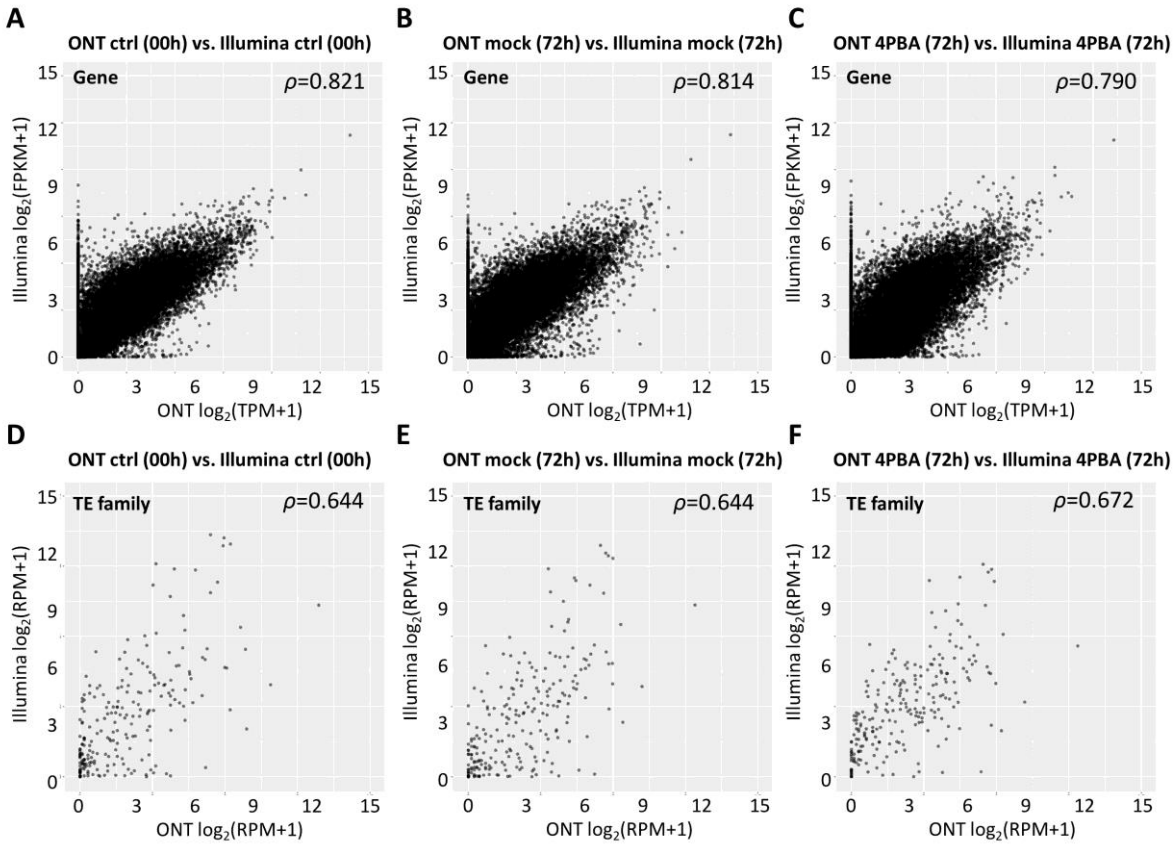



**Figure C.16 Comparisons of gene and TE expression quantified from ONT versus from Illumina Truseq sequencing libraries**

**(A)-(C)** Gene and **(D)-(F)** TE family expression level quantified from ONT was compared with that of Illumina Truseq sequencing data. The gene expression levels were given as transcripts per million mapped reads (TPM) for the ONT libraries (x-axes) and as fragments per kilobase per million mapped reads (FPKM) for the Illumina libraries (y-axes). TE expression levels obtained from both sequencing platforms were given as reads per million mapped reads. Spearman's correlation coefficient ρ was given per comparison, with each point representing gene expression levels.

## C.12 Investigation of autonomous TE loci with the breadth of coverage > 90% across domains necessary for autonomous mobilization

**Table C.13 Investigation of autonomous TE loci with the breadth of coverage > 90% across domains necessary for autonomous mobilization**

| TE type | TE superfamily | # of annotated TEs | # of intact TEs | # of full transcription | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | Ctrl (T=0) | | Mock 72h | | 4PBA 72h | |
| LTR-TEs | Copia | 44,598 | 1,182 | Copia-3:<br>Copia-75: | 3<br>1 | Copia-12: | 7 | Copia-12: | 8 |
| | Gypsy | 64,827 | 1,680 | Gypsy-V1: | 1 | | 0 | Gypay-V1: | 2 |
| | Sum | 109,425 | 1,680 | | 5 | | 7 | | 10 |
| LINE | LINE | 23,477 | 122 | VLINE1:<br>VLINE7:<br>VLINE8: | 1<br>1<br>4 | VLINE8: | 1 | VLINE8: | 3 |
| | Sum | 23,447 | 122 | | 6 | | 1 | | 3 |
| TIR-TEs | hAT | 15,374 | 170 | hAT-7: | 6 | hAT-7: | 6 | hAT-7: | 6 |
| | MULE | 27,336 | 25 | | 0 | | 0 | | 0 |
| | Harbinger | 32,053 | 3 | | 0 | | 0 | | 0 |
| | CACTA | 12,632 | 10 | | 0 | | 0 | | 0 |
| | Sum | 87,395 | 208 | | 6 | | 6 | | 6 |

## C.13 Heatmaps of differentially expressed genes in 4PBA treatment

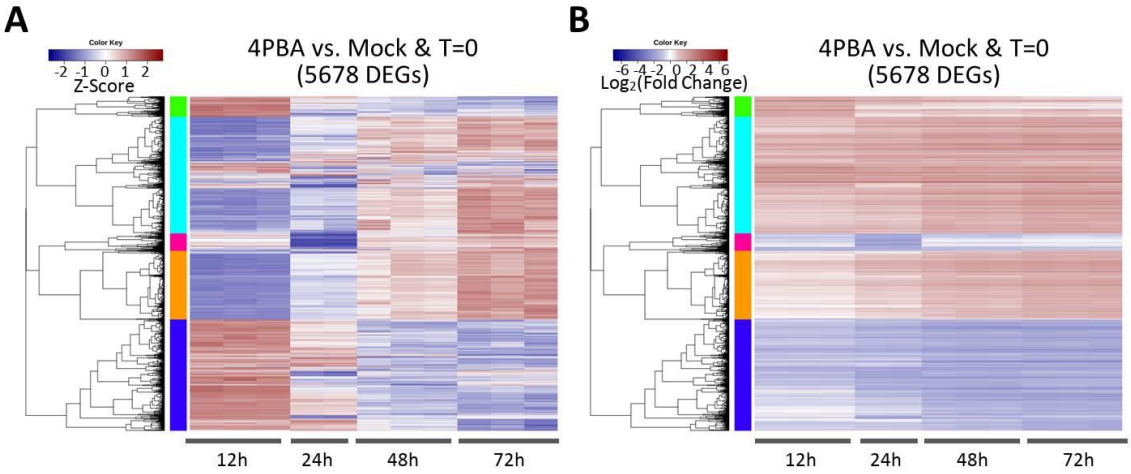



**Figure C.17 Heatmaps of DEGs of 4PBA treatment**

DEGs in 4PBA treatments were illustrated by heatmaps using **(A)** Z-score and **(B)** $\log_2$(fold change).

## C.14 Enriched GO networks of DEGs in 4PBA treatment

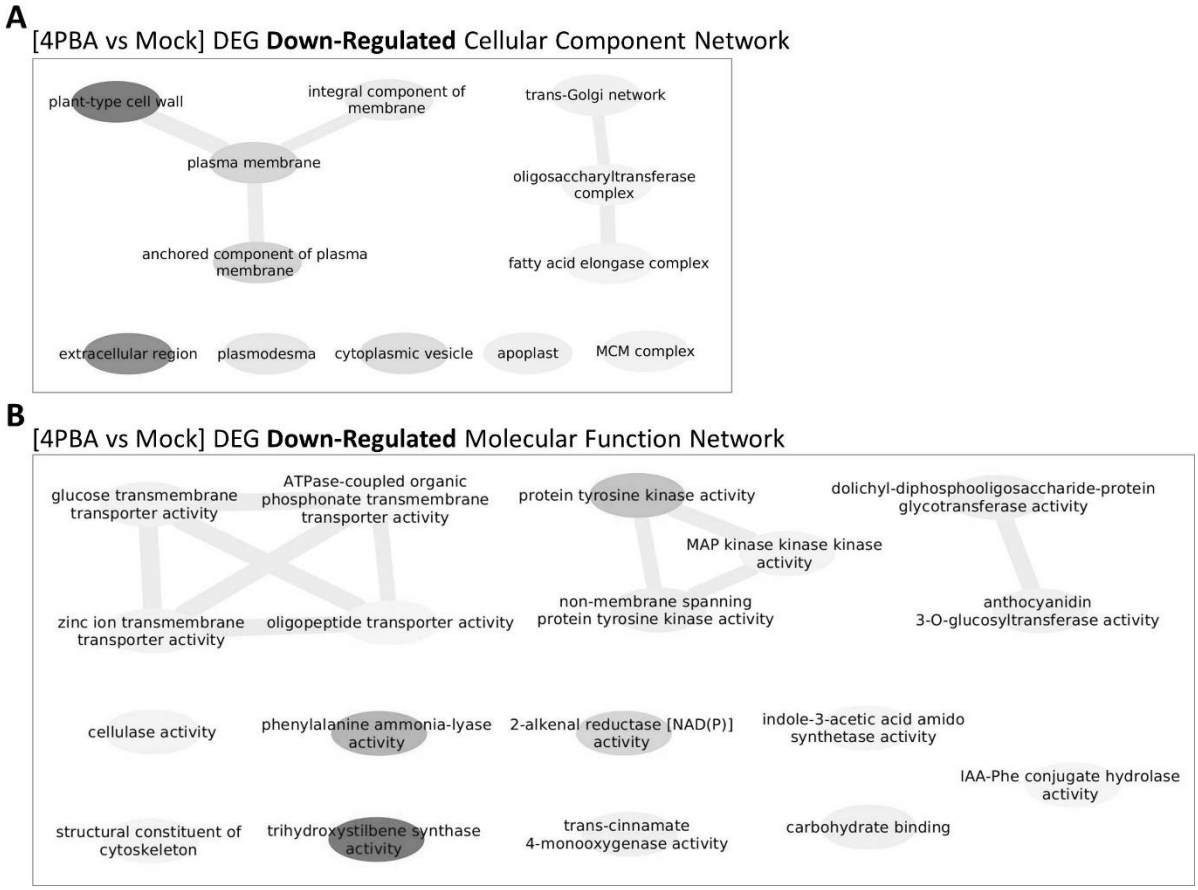



**Figure C.18 Enriched GO networks of down-regulated DEGs in 4PBA treatment**

Significant GO terms ($p < 0.05$) in **(A)** cellular component network and **(B)** molecular function network for down-regulated DEGs in 4PBA treatments. Links denote closely related GO term clusters, among which the darker, the more significantly enriched (lower REVIGO p-value).

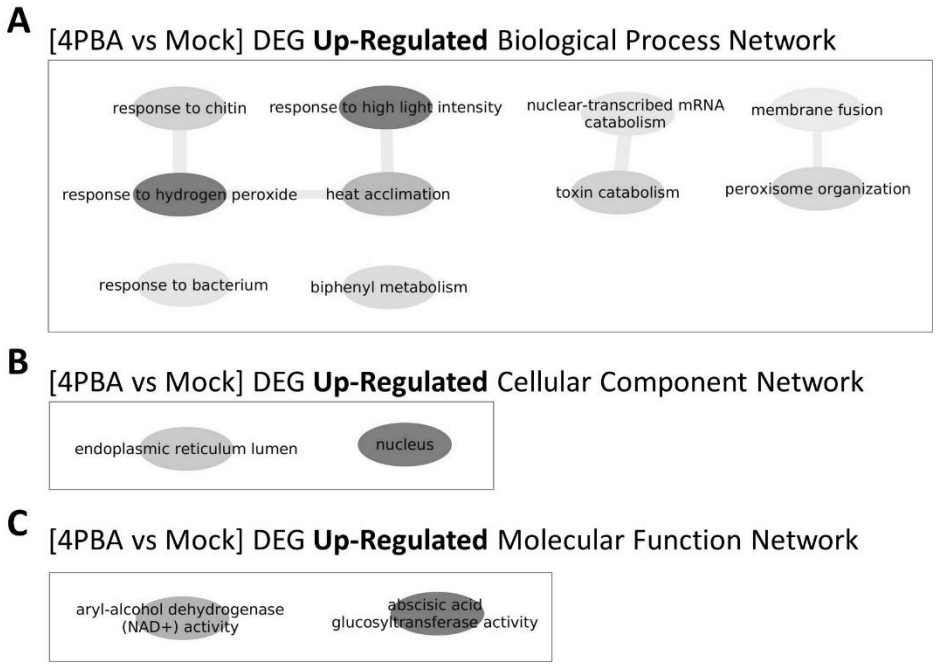



**Figure C.19 Enriched GO networks of up-regulated DEGs in 4PBA treatment**

Significant GO terms ($p < 0.05$) in **(A)** biological process, **(B)** cellular component and **(C)** molecular function networks for up-regulated DEGs in 4PBA treatments. Links denote closely related GO term clusters, among which the darker, the more significantly enriched (lower REVIGO p-value).

## C.15 List of differentially expressed miRNAs

**Table C.14 List of differentially expressed miRNAs in mock treatment**

| **Mock vs. T=0** | | | | |
|---|---|---|---|---|
| | pattern | | Targeted gene or TE (E≤2) | Putative function of targeted gene and notes for targeted TEs |
| vvi-miR3630-3p | Up | vvi-miR3630-3p<br>VIT_207s0005g05420 | 22 CACGUAGUCUCUCUAAGGGUUU 1<br>:: :.:::::::::::::::<br>2938 AAUCACCGGAGAGAUUCCCAAA 2959 | lrr receptor-like serine threonine-protein kinase rch1-like |
| | | vvi-miR3630-3p<br>VIT_218s0072g01230 | 22 CACGUAGUCUCUCUAAGGGUUU 1<br>.::.::::::::..::::<br>3951 AAAUAUUAGAGAGAUUUUCAAA 3972 | tmv resistance protein n-like |
| vvi-miR396b | Up | vvi-miR396b<br>VLINE5 | 20 UCAAGUUCUUUCGACACCUU 1<br>:::: :::::.:::::::::.<br>72 AGUUGAAGAAGGCUGUGGAG 91 | |
| | | vvi-miR396b<br>VLINE2 | 20 UCAAGUUCUUUCGACACCUU 1<br>.::::::: .::::::::::::<br>1639 GGUUCAAAGAAGCUGUGGAA 1658 | |
| | | vvi-miR396b<br>Mutavine-17 | 20 UCAAGUUCUUUCGACACCUU 1<br>::::::::::.::::::: :<br>4662 UGUUCAAGAAGGCUGUGGCA 4681 | |
| | | vvi-miR396b<br>VIT_202s0012g02250 | 20 UCAAGUUCUUUCGACACCUU 1<br>.::::.:::::::.::::::::<br>1863 GGUUUGAGAAAGUUGUGGAA 1882 | transcription factor hbp-1a |
| | | vvi-miR396b<br>VIT_202s0033g01260 | 20 UCAAGUUCUUUCGACACCUU 1<br>::::::::::.::::::: :<br>354 UGUUCAAGAAGGCUGUGGCA 373 | uncharacterized protein |
| | | vvi-miR396b<br>VIT_218s0089g00100 | 20 UCAAGUUCUUUCGACACCUU 1<br>:::: .:.::::::::::::<br>1656 UGUUCCGGGAAGCUGUGGAA 1675 | tmv resistance protein n-like |
| | | vvi-miR396b<br>VIT_218s0089g00100 | 20 UCAAGUUCUUUCGACACCUU 1<br>:::: .:.::::::::::::<br>4404 GGUUUGAGAAAGUUGUGGAA 4423 | |

**Table C.15 List of differentially expressed miRNAs in 4PBA treatment**

| 4PBA vs. mock and T=0 | | | | |
|---|---|---|---|---|
| | pattern | | Gene or TE target (E≤2) | Putative function of targeted gene and notes for targeted TEs |
| new_miR_20 | Down | new_miR_20<br>VIT_202s0025g01760 | 21 CCAGCCUAGCUCAACUUUAAA 1<br>::::::::::::::::::::<br>338 GGUCGGAUCGAGUUGAAAUUU 358 | cellulose synthase-like protein g3-like "ATCSLG1, CSLG1" |
| | | new_miR_20<br>VIT_200s0319g00010 | 21 CCAGCCUAGCUCAACUUUAAA 1<br>:::: ::::::::::::::::<br>288 GGUCCGAUCGAGUUGAAAUUU 308 | NA |
| | | new_miR_20<br>VIT_201s0010g03170 | 21 CCAGCCUAGCUCAACUUUAAA 1<br>::::::::::::::::::::<br>263 GGUCCGAUCGAGUUGAAAUUU 283 | NA |
| | | new_miR_20<br>VIT_213s0019g01300 | 21 CCAGCCUAGCUCAACUUUAAA 1<br>::::::::::::::::::::<br>354 GGUCCGAUCGAGUUGAAAUUU 374 | NA |
| | | new_miR_20<br>VIT_213s0019g01300 | 21 CCAGCCUAGCUCAACUUUAAA 1<br>::::::::::::::::::::<br>534 GGUCCGAUCGAGUUGAAAUUU 554 | |
| | | new_miR_20<br>VIT_203s0097g00240 | 21 CCAGCCUAGCUCAACUUUAAA 1<br>::::::::::::::::::::<br>278 GGUCCGAUCGAGUUGAAAUUU 298 | unnamed protein product |
| | | new_miR_20<br>VIT_203s0097g00240 | 21 CCAGCCUAGCUCAACUUUAAA 1<br>::::::::::::::::::::<br>458 GGUCCGAUCGAGUUGAAAUUU 478 | |
| | | new_miR_20<br>VIT_214s0006g02340 | 21 CCAGCCUAGCUCAACUUUAAA 1<br>::::::::::::::::::::<br>129 GGUCCGAUCGAGUUGAAAUUU 149 | NA |
| vvi-miR171e | Down | vvi-miR171e<br>VIT_204s0023g01380 | 21 CUAUAACCGCGCCGAGUUAGU 1<br>::::::::::::::::::::<br>656 GAUAUUGGCGCGGCUCAAUCA 676 | gras family transcription factor "AtHAM4, HAM4" |
| | | vvi-miR171e<br>VIT_215s0048g00270 | 21 CUAUAACCGCGCCGAGUUAGU 1<br>::::::::::::::::::::<br>1297 GAUAUUGGCGCGGCUCAAUCA 1317 | gras family transcription factor "ATHAM3, HAM3, LOM3, SCL6-IV" |
| | | vvi-miR171e<br>VIT_202s0154g00400 | 21 CUAUAACCGCGCCGAGUUAGU 1<br>::::::::::::::::::::<br>1724 GAUAUUGGCGCGGCUCAAUCA 1744 | gras family transcription factor "ATHAM3, HAM3, LOM3, SCL6-IV" |
| | | vvi-miR171e<br>VIT_214s0068g01780 | 21 CUAUAACCGCGCCGAGUUAGU 1<br>::::::::: ::::::::::<br>2805 GAUAUUGGCACGGCUCAAUCA 2825 | squamosa promoter-binding-like protein 16-like "SPL13, SPL13B" |
| vvi-miR171f | Down | vvi-miR171f<br>VIT_215s0048g00270 | 21 UCACUAUAACCGCGCCGAGUU 1<br>:: ::::::::::::::::::<br>1294 AGGGAUAUUGGCGCGGCUCAA 1314 | gras family transcription factor factor "ATHAM3, HAM3, LOM3, SCL6-IV" |
| | | vvi-miR171f<br>VIT_202s0154g00400 | 21 UCACUAUAACCGCGCCGAGUU 1<br>:: ::::::::::::::::::<br>1721 AGGGAUAUUGGCGCGGCUCAA 1741 | gras family transcription factor "ATHAM3, HAM3, LOM3, SCL6-IV" |
| | | vvi-miR171f<br>VIT_204s0023g01380 | 21 UCACUAUAACCGCGCCGAGUU 1<br>.: ::::::::::::::::::<br>653 GGCGAUAUUGGCGCGGCUCAA 673 | gras family transcription factor "AtHAM4, HAM4" |
| | | vvi-miR171f<br>VIT_214s0068g01780 | 21 UCACUAUAACCGCGCCGAGUU 1<br>:::::::::: ::::::::<br>2802 CAUGAUAUUGGCACGGCUCAA 2822 | squamosa promoter-binding-like protein 16-like "SPL13, SPL13B" |
| | | vvi-miR171f<br>VIT_218s0001g03310 | 21 UCACUAUAACCGCGCCGAGUU 1<br>.:::::::::::. ::::::::<br>375 GGUGAUAUUGGUUCGGCUCAA 395 | gras family protein |
| | | vvi-miR171f<br>VIT_212s0059g00650 | 21 UCACUAUAACCGCGCCGAGUU 1<br>: :::::::::.::::.::::<br>71 CGGGAUAUUGGUGCGGUUCAA 91 | NA |
| vvi-miR3633a-5p | Down | vvi-miR3633a-5p<br>VIT_201s0026g02650 | 20 GAGAGGAUUGGUAGGUAAGG 1<br>:: :::::::::::::::..<br>4900 CUGUCCUAACCAUCCAUUUU 4919 | kh domain-containing protein at4g18375-like |
| | | vvi-miR3633a-5p<br>VIT_219s0015g02930 | 20 GAGAGGAUUGGUAGGUAAGG 1<br>:::.::::::.:::..:::::<br>49 GUCUUCUAACUAUUUAUUCC 68 | unnamed protein product |
| vvi-miR3635-5p | Down | vvi-miR3635-5p<br>VIT_219s0015g00020 | 21 GAUAAUACGGGGUGUGUACGG 1<br>::::::::::::::::::<br>750 CCGCUAUGCCCCACACAUGCC 772 | abc transporter c family member 3-like "ABCC3, ATMRP3, MRP3" |
| | | vvi-miR3635-5p<br>VIT_219s0015g00010 | 21 GAUAAUACGGGGUGUGUACGG 1<br>::::::::::::::::::<br>432 CCGCUAUGCCCCACACAUGCC 452 | abc transporter c family member 3-like "ABCC3, ATMRP3, MRP3" |
| | | vvi-miR3635-5p<br>VIT_219s0015g00060 | 21 GAUAAUACGGGGUGUGUACGG 1<br>::::::::::::::::::<br>432 CCGCUAUGCCCCACACAUGCC 452 | multidrug resistance protein abc transporter family "ABCC3, ATMRP3, MRP3" |
| | | vvi-miR3635-5p | 21 GAUAAUACGGGGUGUGUACGG 1<br>:::::::::::::::::: | abc transporter c family member 3-like |

| | | | | |
|---|---|---|---|---|
| | | VIT_219s0015g00050 | 773 CCGCUAUGCCCCACACAUGCC 793 | "ABCC3, ATMRP3, MRP3" |
| vvi-miR3636-5p | Down | vvi-miR3636-5p<br>VIT_203s0017g01370 | 24 CUUAGAUAGUUUCUUCGUUUGGCU 1<br>1921 GCUGUUGUCAAAGAAGUAGACUGA 1944 | NA |
| | | vvi-miR3636-5p<br>VIT_211s0016g00420 | 24 CUUAGAUAGUUUCUUCGUUUGGCU 1<br>1327 UGGAAUGUCAAAGAAUCAAACCGA 1350 | splicing factor SMU2 |
| vvi-miR3640-5p | Down | NA | | |
| vvi-miR390 | Down | vvi-miR3636-5p<br>VIT_212s0059g01410 | 21 CCGCGAUAGGGAGGACUCGAA 1<br>200 GGCGAUAUCUCUCCUGAGCUU 220 | NA |
| | | vvi-miR3636-5p<br>VIT_210s0003g01890 | 21 CCGCGAUAGGGAGGACUCGAA 1<br>319 GGCGUUCUUCCUCCUGAGCUU 339 | lrr receptor-like serine threonine-protein kinase rfk1 |
| vvi-miR396d | Down | vvi-miR396d<br>VLINE5 | 21 GUCAAGUUCUUUCGACACCUU 1<br>71 CAGUUGAAGAAGGCUGUGGAG 91 | |
| | | vvi-miR396d<br>VLINE2 | 21 GUCAAGUUCUUUCGACACCUU 1<br>1638 UGGUUCAAAGAAGCUGUGGAA 1658 | |
| | | vvi-miR396d<br>Mutavine-17 | 21 GUCAAGUUCUUUCGACACCUU 1<br>4661 GUGUUCAAGAAGGCUGUGGCA 4681 | |
| | | vvi-miR396d<br>VIT_202s0012g02250 | 21 GUCAAGUUCUUUCGACACCUU 1<br>1862 UGGUUUGAGAAAGUUGUGGAA 1882 | transcription factor hbp-1a |
| | | vvi-miR396d<br>VIT_202s0033g01260 | 21 GUCAAGUUCUUUCGACACCUU 1<br>353 GUGUUCAAGAAGGCUGUGGCA 373 | uncharacterized protein |
| | | vvi-miR396d<br>VIT_218s0089g00100 | 21 GUCAAGUUCUUUCGACACCUU 1<br>1655 UUGUUCCGGGAAGCUGUGGAA 1675 | tmv resistance protein n-like |
| | | vvi-miR396d<br>VIT_218s0089g00100 | 21 GUCAAGUUCUUUCGACACCUU 1<br>4403 UUGUUCCGGGAAGCUGUGGAA 4423 | |
| vvi-miR398b | Up | NA | | |
| vvi-miR408 | Up | vvi-miR408<br>VIT_207s0005g02730 | 21 CGGUCCCUUCUCCGUCACGUA 1<br>1650 GCCAGGGAAGAGGCAGUGCAU 1670 | NA |
| | | vvi-miR408<br>VIT_218s0001g15240 | 21 CGGUCCCUUCUCCGUCACGUA 1<br>790 GUGAGGGAAGAGGCAGUGCAG 810 | basic blue protein ARPN. Encodes plantacyanin one of blue copper proteins. Involved in anther development and pollination. |

## C.16 List of TE families producing siRNA at a level > 100 RPM

### Table C.16 List of TE families producing siRNA at a level > 100 RPM

| Rank | TE family | Mean RPM | Rank | TE family | Mean RPM | Rank | TE family | Mean RPM |
|---|---|---|---|---|---|---|---|---|
| 1 | Gypsy-GYVIT1 | 1069 | 34 | Gypsy-21 | 248 | 66 | Copia-34 | 138 |
| 2 | Gypsy-16 | 921 | 35 | Gypsy-11 | 248 | 67 | Gypsy-33 | 136 |
| 3 | Copia-31 | 699 | 36 | MULE-MuDR-8 | 244 | 68 | Copia-Tvv1 | 135 |
| 4 | Gypsy-3 | 563 | 37 | MULE-MuDR-4 | 233 | 69 | MULE-Mutavine-14 | 134 |
| 5 | MULE-Jitvine-2 | 550 | 38 | hAT-VIHAT1 | 227 | 70 | Copia-58 | 132 |
| 6 | Gypsy-Gret1 | 529 | 39 | hAT-VIHAT3 | 225 | 71 | Copia-40 | 129 |
| 7 | Gypsy-26 | 527 | 40 | Copia-47 | 219 | 72 | Copia-75 | 128 |
| 8 | hAT-13 | 497 | 41 | Copia-44 | 218 | 73 | Copia-9 | 125 |
| 9 | Copia-88 | 441 | 42 | Gypsy-17 | 216 | 74 | Gypsy-2 | 124 |
| 10 | Copia-3 | 420 | 43 | Gypsy-7 | 215 | 75 | Gypsy-V1 | 123 |
| 11 | Copia-15 | 415 | 44 | Copia-11 | 214 | 76 | Copia-26 | 123 |
| 12 | Gypsy-23 | 409 | 45 | CACTA-13 | 212 | 77 | CACTA-6 | 121 |
| 13 | Gypsy-12 | 408 | 46 | Gypsy-29 | 210 | 78 | CACTA-4N1 | 119 |
| 14 | Copia-33 | 397 | 47 | MULE-MuDR-6 | 209 | 79 | Copia-49 | 118 |
| 15 | Gypsy-14 | 377 | 48 | Copia-76 | 208 | 80 | CACTA-5 | 118 |
| 16 | Copia-10 | 370 | 49 | Copia-32 | 203 | 81 | Gypsy-19 | 116 |
| 17 | Caulimovirus-CAULIV11 | 359 | 50 | Caulimovirus-1 | 201 | 82 | Copia-46 | 116 |
| 18 | hAT-7 | 359 | 51 | Gypsy-20 | 197 | 83 | Copia-85 | 115 |
| 19 | Gypsy-27 | 358 | 52 | hAT-10 | 194 | 84 | MULE-MuDR-12 | 111 |
| 20 | MULE-MuDR-21 | 356 | 53 | Gypsy-9 | 189 | 85 | MULE-MuDR-7 | 109 |
| 21 | MULE-Mutavine-17 | 353 | 54 | MULE-MuDR-9 | 187 | 86 | Harbinger-VHARB4 | 101 |
| 22 | Gypsy-18 | 349 | 55 | Caulimovirus-2 | 185 | 87 | Gypsy-22 | 100 |
| 23 | MULE-Mutavine-18 | 346 | 56 | Copia-17 | 183 | 88 | CACTA-N3 | 99 |
| 24 | MULE-MuDR-18 | 344 | 57 | Copia-70 | 174 | 89 | Copia-5 | 98 |
| 25 | Copia-22 | 330 | 58 | CACTA-7 | 164 | 90 | Copia-28 | 93 |
| 26 | MULE-MUDRAVI1 | 315 | 59 | Gypsy-4 | 157 | 91 | CACTA-3 | 93 |
| 27 | Copia-23 | 305 | 60 | hAT-VIHAT2 | 155 | 92 | Gypsy-34 | 92 |
| 28 | Gypsy-6 | 297 | 61 | CACTA-1 | 154 | 93 | Copia-59 | 89 |
| 29 | MULE-MuDR-22 | 295 | 62 | Copia-94 | 153 | 94 | LINE-VLINE2 | 82 |
| 30 | Copia-89 | 285 | 63 | CACTA-4 | 139 | 95 | Harbinger-VHARB-N2 | 82 |
| 31 | Gypsy-13 | 280 | 64 | CACTA-2 | 139 | 96 | Copia-86 | 79 |
| 32 | MULE-MuDR-13 | 256 | 65 | hAT-12 | 138 | 97 | LINE-VLINE8 | 79 |
| 33 | MULE-MUDRAVI2 | 254 | | | | | | |

# Appendix D

# Computational scripts

All content of Appendix D is available at https://figshare.com/s/248be5bfa9ecf6471e08 .

## D.1 RNAseq analysis pipeline for the identification of TE expression candidates

## D.2 Analysis scripts of the characteristics of TE expression candidates

## D.3 Analysis scripts of the transcriptional relationship between TEs and co-localized genes

## D.4 Analysis scripts of ONT cDNA sequencing data

## D.5 Analysis scripts of sRNA sequencing data